



IMPROVING CAPE CANAVERAL'S NEXT-DAY
THUNDERSTORM FORECASTING USING
A MESO-ETA MODEL-BASED INDEX

THESIS

John C. Crane, First Lieutenant, USAF
AFIT/GM/ENP/99M-4

19990402 010

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

Approved for public release; distribution unlimited.

DTIC QUALITY INSPECTED 2

IMPROVING CAPE CANAVERAL'S NEXT-DAY THUNDERSTORM
FORECASTING USING A MESO-ETA MODEL-BASED INDEX

THESIS

Presented to the Faculty of the Graduate School of Engineering
of the Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the
Requirements for the Degree of
Master of Science in Meteorology

John C. Crane, B.S.

First Lieutenant, USAF

March, 1999


Approved for public release; distribution unlimited

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense of the U. S. Government.

IMPROVING CAPE CANAVERAL'S NEXT-DAY THUNDERSTORM
FORECASTING USING A MESO-ETA MODEL-BASED INDEX

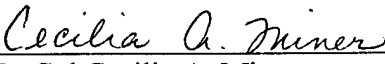
John C. Crane, B.S.
First Lieutenant, USAF

Approved:



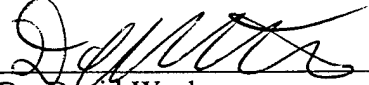
Lt Col Michael K. Walters
Chair, Advisory Committee

1 Mar 99
Date



Lt Col Cecilia A. Miner
Member, Advisory Committee

2 Mar 99
Date



Dr. David Weeks
Member, Advisory Committee

2 Mar 99
Date

Acknowledgements

Many people have been instrumental in the final product of this thesis. The first person I would like to thank is Lieutenant Colonel Michael K. Walters, my thesis advisor, for his guidance and trust in me to put out a worthy thesis. I would also like to thank the other members of my thesis committee for taking the time to see that I get the most out of my thesis experience.

For guidance in statistical matters, I would like to thank Mr. Dan Reynolds for always being there to put me in the right direction. His knowledge of statistical software packages and their use helped me through many tough hours of regression analysis. He also aided immeasurably by directing me to references of true value.

I would also like to thank Mr. William Roeder, my thesis sponsor, Chief Staff Meteorologist of the 45th Weather Squadron at Patrick AFB and Cape Canaveral Air Station, Florida. His insight to important physical processes at the Cape aided in my search for important variables to include in my model. Throughout the course of my thesis work I received many a long e-mail packed with a wealth of valuable information. Mr. Roeder was also was beneficial in my search for quality references to help with this thesis.

Last, I would like to thank someone that might better be left unnamed for giving me general hope in life, the Dodge gym for giving me a place to maintain my sanity, and my computer for not belching during this arduous task.

John C. Crane

Table of Contents

Acknowledgements	iii
Table of Contents	iv
List of Figures	vii
List of Tables	viii
Abstract	ix
1. Introduction	1
1.1 Overview	1
1.2 Background	3
1.3 Problem Statement	5
1.4 Importance of Research	5
1.5 Overall Approach	7
1.6 Organizational Overview	8
2. Literature Review and Theory	10
2.1 Thunderstorm Fundamentals	10
2.2 Numerical Modeling	13
2.2.1 Introduction	13
2.2.2 History	14
2.2.3 Modeling Fundamentals	15
2.2.3.1 Numerical Method Used	16
2.2.3.2 Scale of Interest	19
2.2.3.3 Equations Used	19

2.2.3.4 Initial and Boundary Conditions.....	20
2.2.4 The Mesoscale Eta Model	21
2.2.5 Previous Work	28
2.2.5.1 The NPTI.....	29
2.2.5.2 Other Studies	30
3. Methodology	32
3.1 Overview	32
3.2 Data Processing	32
3.3 Statistical Methods	40
3.3.1 Logistic Regression	40
3.3.2 Verification	43
4. Analysis and Results	50
4.1 Overview	50
4.2 Variables Found to be Important	50
4.3 Significance of Contingency Tables	57
4.4 Verification Statistics	57
5. Summary and Recommendations.....	60
5.1 Overview	60
5.2 Summary.....	60
5.3 Recommendations	61
Appendix A: Variables Considered	63
Appendix B: Percent Deviance Explained and T-Value Tables.....	65
Appendix C: Verification Statistics	67

Appendix D: Mathcad Template for Calculating NPTI.....	72
Appendix E: Mathcad program for 99% Ellipse Winds.....	75
Bibliography	81
Vita.....	85

List of Figures

Figure 1. Map of Florida showing location of Cape Canaveral.....	2
Figure 2. Cape Canaveral, Florida	4
Figure 3. Portion of Seven-Day Planning Forecast	6
Figure 4. Map showing the 29-km, 32-km, and 48-km eta domains	23
Figure 5. 45 Vertical Levels of 32-km eta.....	24
Figure 6. Semi-Staggered Arakawa E-grid.....	25
Figure 7. Two-day Persistence.....	38
Figure 8. 2X2 Contingency Table.....	45

List of Tables

Table 1. Variables Found to be Important to Thunderstorm Occurrence.....	51
Table 2. Beta Coefficients for T1T2.....	51
Table 3. Beta Coefficients for T2T3.....	52
Table 4. Beta Coefficients for Full Set.....	52
Table B-1. T1T2-A: Percent Deviance Explained and T-Values.....	65
Table B-2. T1T2-B: Percent Deviance Explained and T-Values.....	65
Table B-3. T1T2-C: Percent Deviance Explained and T-Values.....	65
Table B-4. T2T3-A: Percent Deviance Explained and T-Values.....	65
Table B-5. T2T3-B: Percent Deviance Explained and T-Values.....	66
Table B-6. T2T3-C: Percent Deviance Explained and T-Values.....	66
Table B-7. FULL-A: Percent Deviance Explained and T-Values.....	66
Table B-8. FULL-B: Percent Deviance Explained and T-Values.....	66
Table B-9. FULL-C: Percent Deviance Explained and T-Values.....	66
Table C-1. Verification Statistics for T1T2-A.....	67
Table C-2. Verification Statistics for T1T2-B.....	67
Table C-3. Verification Statistics for T1T2-C.....	68
Table C-4. Verification Statistics for T2T3-A.....	68
Table C-5. Verification Statistics for T2T3-B.....	69
Table C-6. Verification Statistics for T2T3-C.....	69
Table C-7. Verification Statistics for FULL-A.....	70
Table C-8. Verification Statistics for FULL-B.....	70
Table C-9. Verification Statistics for FULL-C.....	70

Abstract

Reliable thunderstorm forecasts are essential to safety and resource protection at Cape Canaveral. Current methods of forecasting day-2 thunderstorms provide little improvement over forecasting by persistence alone and are therefore in need of replacement. This thesis focuses on using the mesoscale eta model to develop an index for improved forecasting of day-2 thunderstorms.

Surface observations from the shuttle landing facility and the coincident output of the mesoscale eta forecast model were collected for the period of 1 May to 14 Sep 1998. Variables extracted from the eta forecast model output, as well as derived variables that incorporate the eta output variables, were divided into three data sets. A univariate logistic regression with the occurrence of a thunderstorm in the surface observation (the "truth") as the dependent variable, and the output/derived values from the eta model as the independent variable, discarded all but 94 of over 250 predictors that were considered important to thunderstorm occurrence. The data sets were further divided into model-building and validation sets for the purpose of building a logistic regression model. Regression coefficients were developed using the model-building sets and then applied to the validation sets for the purpose of forecasting a thunderstorm. Verification of the forecasts was accomplished using standard accuracy and skill measures, and comparisons were made against persistence and against model-based forecasts of the Neumann-Pfeffer Thunderstorm Index (NPTI).

In cases where both persistence and the index developed through this research (called the eta thunderstorm index or ETI) were found to be significant, the ETI consistently outperformed persistence. Due to the small sample size of this research, further study is necessary to validate the results of this thesis.

IMPROVING CAPE CANAVERAL'S NEXT-DAY THUNDERSTORM FORECASTING USING A MESO-ETA MODEL-BASED INDEX

1. Introduction

1.1 Overview

Cape Canaveral Air Station and the Kennedy Space Center (CCAS/KSC) are an important part of America's space program. With nearly thirty percent of all space launch attempts being scrubbed or delayed due to weather (Roeder, 1998), it is easy to see the importance placed on forecasting launch weather. The 45th Weather Squadron (45 WS) based at Patrick Air Force Base, Florida is responsible for this forecast challenge. In their quest to visualize their mission statement, "exploit the weather to assure access to air and space," they sponsor operational research such as this thesis. By focusing research efforts on the weather phenomena that cause the greatest detriment to launch operations, they seek to minimize the impact of the weather.

The most salient weather condition to launch planning, launch operations, and even day-to-day ground operations during the warm season (May 1-Sep 30) is thunderstorms. Therefore, correctly forecasting thunderstorm occurrence is of primary importance in reducing the percentage of scrubbed or delayed launches (Bauman and Businger, 1996). Current forecast methods provide only a marginal improvement over persistence (the probability of having a thunderstorm today if there was one yesterday) when forecasting next-day thunderstorms for planning purposes and therefore are in need

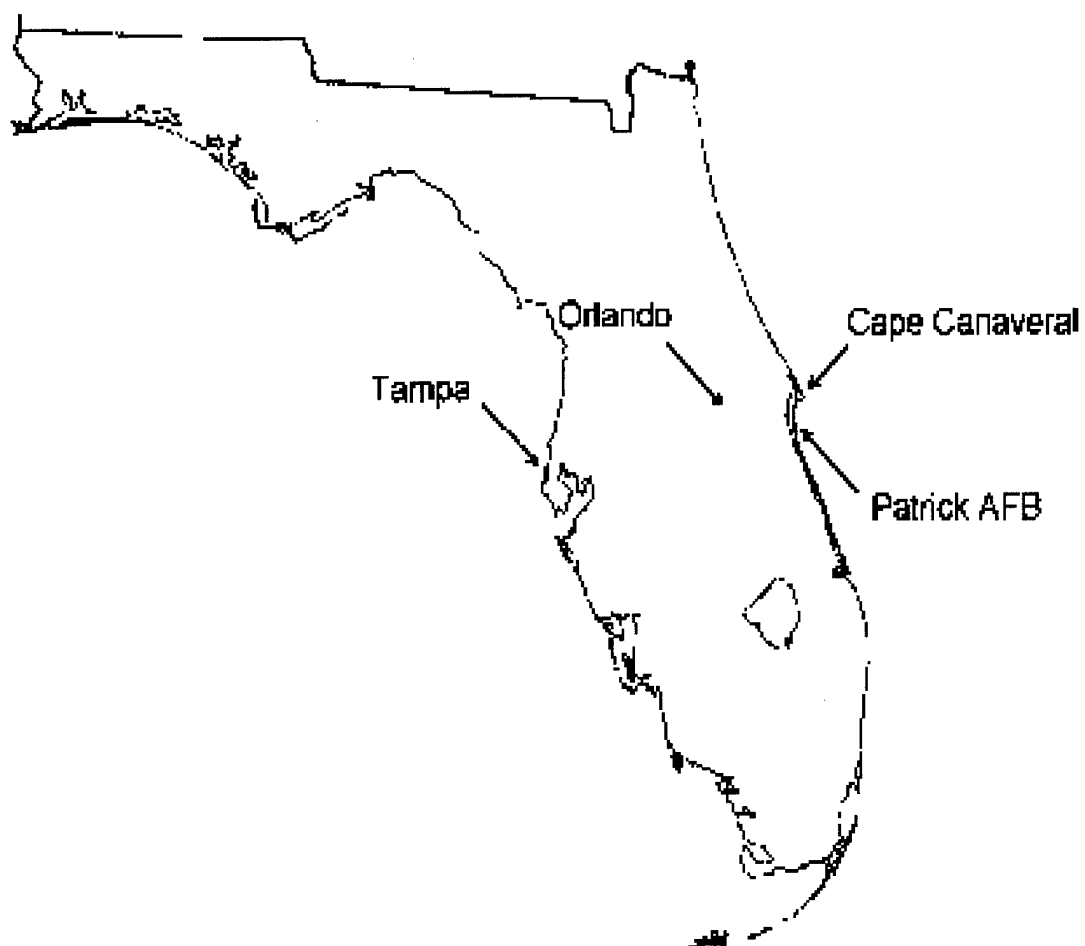


Figure 1 Map of Florida showing location of Cape Canaveral

of replacement. By exploiting current knowledge of the weather through meteorological models and statistics, an improved thunderstorm forecast method can be attained. In an attempt to find this forecast method, it is necessary to analyze which factors are pertinent to thunderstorm formation and, therefore, forecasting, near Cape Canaveral.

1.2 Background

Forecasting thunderstorms at Cape Canaveral has always presented quite a challenge due to the complexity of boundary layer interactions in that area (Zhong and Takle, 1993). By definition, a cape is a piece of land that projects into a body of water. This projection in itself causes enough irregularity in the sea/land breeze pattern to present a forecast challenge (see Figure 1 and Figure 2). On top of this, the problem is compounded by a plethora of other boundary interactions such as two sea breezes (east and west coast of Florida), two river breezes (from the Indian and Banana rivers), convective outflows (even from the previous day), lake breezes, cloud shadow breezes, and even soil moisture breezes (Roeder, 1998). With all these complicating factors, it is easy to see why thunderstorms are difficult to forecast near Cape Canaveral. In the past, a variety of tools have been formulated to help the forecaster determine the probability of thunderstorm occurrence.

Currently, forecasters at the 45 WS are using a 30-year-old technique as their primary objective tool for forecasting the thunderstorm probability for the current day (Roeder, 1998). This tool, called the Neumann-Pfeffer Thunderstorm Index (NPTI), was designed specifically for use at Cape Canaveral (Neumann, 1971). The NPTI uses variables obtained from the current 1200Z atmospheric rawinsonde sounding to give a

percent probability of thunderstorm occurrence for that day. To forecast the day-2 through day-7 thunderstorm probability, forecasters use a more subjective approach of looking at model guidance and using their meteorological knowledge (Roeder, 1998).

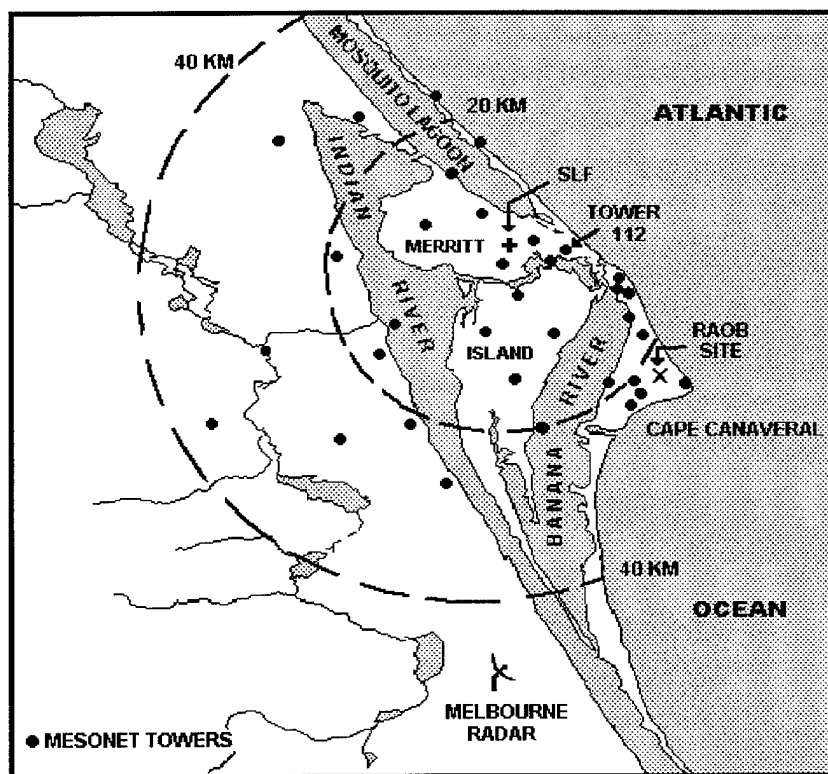


Figure 2. Cape Canaveral, Florida

Over the years, it has been observed that the probability of precipitation at Cape Canaveral relates well to the probability of thunderstorms (since most precipitation in the warm season is convective), and it is therefore one of the primary indicators that forecasters use when forecasting day-2 through day-7 thunderstorm probability (Roeder,

1998). The 45th WS puts out a daily seven-day planning forecast that shows the probability of thunderstorms on station for each of the seven days. This seven-day forecast is one of the major tools that planners at the Cape use to schedule pre-launch space operations. The goal of this thesis is to improve the day-2 lightning probability (see Figure 3) through the development of a model-based index.

1.3 Problem Statement

Can the eta model output from the National Centers for Environmental Prediction (NCEP) be used to develop a statistically significant model-based thunderstorm index that is tailored for Cape Canaveral? If so, how does the index perform against persistence and against NPTI forecasts based on the eta model output?

1.4 Importance of Research

As mentioned above, the 45th WS is tasked daily with providing a forecast of the probability of thunderstorm occurrence for the current and following six days. This information is used by planners to adjust the ground operations schedules to give the highest probability of the operation being successfully accomplished (Wohlwend, 1998). The index developed by this research aims to improve the day-2 thunderstorm forecast and therefore seeks to minimize wasted resources due to inaccurate thunderstorm forecasting. Additionally, this research can be a foundation on which to build an index that extends past day-2 (mesoscale-eta model only forecasts out to 33 hours) by using a

Seven Day Planning Forecast						
Cape Canaveral Air Station & Kennedy Space Center						
Day 1 Forecast Valid 0700-2400L				Issued:		
		Wednesday	Thursday	Friday	Saturday	Sunday
Forecast		11 Mar 98	12 Mar 98	13 Mar 98	14 Mar 98	15 Mar 98
Sky Condition	AM	Clear	Partly Cloudy	Mostly Cloudy	Partly Cloudy	Partly Cloudy
	PM					
Weather	AM	None	TRW/RW	TRW/RW	None	None
	PM					
Precipitation Probability	AM	0%	60%	Improve this number!		0%
	PM					
Lightning Probability	AM	0%	40%	40%	0%	0%
	PM					
Prevailing Winds	AM	NW 15-20	SW 10-15	SW 10-15	S 10	S 5-10
	PM					

Figure 3 Portion of Seven-Day Planning Forecast

different model such as the Medium Range Forecast (MRF), which forecasts out to 10 days. Lastly, the results may show that the physics of the mesoscale eta model are not yet representative enough of actual processes in the atmosphere to resolve thunderstorms at the scale of those developing in the environment around the Cape.

1.5 Overall Approach

This thesis consisted of five main steps with quality control steps in each process. These five steps are collecting data, processing data, selecting variables, running the regression to find a model, and performing statistics to verify the validity of the model. The original intent of this study was to use model data from 1995 to 1998. However, due to a long series of data-gathering problems, only the model data from 1998 were sufficient for use. Model data for 1998 were obtained from the Environmental Modeling Center (EMC) at NCEP. The data files consisted of one thirteen-megabyte file for each of the 114 days that model data were available. Missing days are attributed to model maintenance and the occasional re-allocation of NCEP computer resources to concentrate forecasting efforts on hurricane movements (Rogers, 1998). Surface observations for KTTS (station identifier for Cape Canaveral, located at SLF in Figure 2) were obtained from the Air Force Combat Climatology Center in Asheville, NC.

Data processing involved transforming the binary gridded model data sent by NCEP to its final form of one text file in matrix format with dimensions 114 by 241. This was accomplished through the use of GEMPAK and multiple Unix scripts, Fortran programs, and Mathcad programs. Another Fortran program was written to find which days, with respect to local time, in the surface observation data had an observed

thunderstorm. For analysis purposes, the days with thunderstorms were then added to the data matrix so that the observations were in the same row as the forecast for that day. Next, the variables in the data file were saved by month and into three blocks of 45 days each.

The variable selection phase consisted of running 1500+ individual univariate logistic regressions on the data files segmented by month. Variables that were found to be statistically important based on the univariate regressions in at least three of the five months were retained in a final data set for analysis. To prepare for the regression step, subsets of the 45-day blocks were taken to provide a model-building and verification set of sufficient sample size for each block. The regression step involved performing a stepwise logistic regression on each of the three model-building sets to determine the regression coefficients. With the coefficients in hand, the last step was to assess the fit of the model, and perform statistics to determine the forecast skill of the regression model. The measures of forecast skill used will be discussed in chapter three.

1.6 Organizational Overview

Chapter two contains a literature review and a discussion of theory involved in this thesis. The chapter begins with a discussion of fundamental thunderstorm theory. Next, the basics of numerical modeling are presented. Following the information on models, the eta model is discussed as well as some evaluations that have been performed on the model's representativeness. The chapter ends with information on previous attempts to improve upon current thunderstorm forecasting techniques to include the theory and a brief history of the NPTI.

Chapter three describes the methodology used through all phases in the preparation of this thesis including the statistical process used along the way.

Chapter four presents the analysis of the data and discusses the results and conclusions that can be drawn from these results.

Chapter five is a summary of the research done and recommendations for future research.

2. Literature Review and Theory

2.1 Thunderstorm Fundamentals

The Florida peninsula observes more thunderstorms than any other location in the United States (U.S. Weather Bureau, 1952). This maximum of thunderstorm activity can be attributed to the virtually ever-present conditions that favor thunderstorm development. These conditions are a moist layer near the ground surface, convective instability, and an initiator or trigger (Ray, 1986). Each of these factors is discussed below.

The proximity of Cape Canaveral to the Atlantic Ocean, to the Gulf of Mexico, and to the vast network of rivers and lakes in the area provides more than ample low-level moisture for thunderstorm development. Low-level moisture availability is easily ascertained by analyzing surface dewpoints and the mean layer relative humidity in the lowest 5000 feet (Ray, 1986). Increases in soil moisture from previous thunderstorms may also contribute to further thunderstorm development. With these numerous moisture sources, the availability of moisture around the Cape is usually a given.

Convective instability is present when within a lifted layer the rate of change of equivalent potential temperature with height is less than zero (Wallace and Hobbs, 1977). The term convective instability, also called potential instability, originated from Carl-Gustav Rossby in the 1930's when it was customary to plot soundings on a Rossby diagram which had equivalent potential temperature as a vertical coordinate (AWS/TR-79/006). The Rossby diagram never gained popular acceptance, and meteorologists in the U.S. moved to using the skew-T Log-p as the primary thermodynamic sounding chart.

Due to the release times of rawinsondes in the U.S., upper air data is usually at least a few hours old when convection starts to develop (Ray, 1986). Also, since virtually all of the tropics are convectively unstable up to about 6 km (Wallace and Hobbs, 1977) other methods for determining the possible areas of thunderstorm development are needed. One method to monitor low-level stability is to lift the surface temperature moist adiabatically to 500 millibars (mb). Larger positive values for the difference between the parcel temperature and the 500-mb temperature indicate decreasing stability. Other methods include monitoring cloud bases, pilot reports, and inversion cap strengths.

A trigger, the third ingredient in thunderstorm development, is anything that causes air to rise. Fronts, jet streams, convergent zones, surface heating, surface troughs, outflow boundaries, and cloud boundaries can all act as triggers. During the summer, moisture and instability are typically always present around the Cape; therefore, the forecasting question becomes not if a thunderstorm will develop, but when and where one will develop (Roeder, 1998). This often is answered based on the location of the triggering mechanism. In the summer, surface heating usually reaches a maximum between 1400 and 1600 local standard time (LST) each day (Bluestein 1993, Vol. II) which coincides with the maximum land/sea temperature difference and the strongest surface winds (Zhong and Takle, 1992). Thus, there may be interaction of two triggering mechanisms in the development of thunderstorms. Cloud boundaries also present an interesting aspect in thunderstorm development. If there is a persistent cloud layer adjacent to a clear area, the surface below will undergo differential heating. This differential heating can cause convergence at the surface which, in turn, leads to rising

air. With the availability of moisture, and the virtually ever-present instability near the Cape, the main forecasting problem is therefore the forecasting of the trigger.

The vast majority of thunderstorms that occur in the warm season of the Florida peninsula are referred to as "air-mass thunderstorms" in order to distinguish them from thunderstorms that occur due to synoptic-scale disturbances (Wallace and Hobbs, 1977). Due to the complex internal structure of cumulus clouds (Holton, 1992) an idealized life cycle of air-mass thunderstorms was developed from a program known as the Thunderstorm Project in the 1940's. This life cycle consists of three stages: cumulus, mature, and dissipating. Air-mass thunderstorms typically have one or more cells that follow this life cycle each cell growing and decaying in rapid succession with the lifetime of an individual cell being about a half hour (Wallace and Hobbs, 1977).

Apart from the life cycle, thunderstorms are generally classified into one of three primary types. These are the single-cell, the multi-cell, and the super-cell. The distinguishing factor that has the most influence on thunderstorm type is the vertical wind shear in the lower troposphere (Holton, 1992) with lower values of shear (<10 m/s below 4 km) indicative of single-cell thunderstorms. Therefore, the weakly sheared low-level environment around the Cape is ripe for single-cell development.

Though these single-cell, air-mass type thunderstorms around the Cape usually don't reach severe criteria, they nonetheless affect daily operations. In order to minimize any impact that thunderstorms have on operations, an ultra-conservative stance has been taken on weather limits for various criteria. These criteria are ceiling and visibility, precipitation, lightning, and winds. Each of the criteria listed below will be discussed in

relation to a space shuttle landing since the space shuttle is essentially a glider (i.e. it has one chance to land) while in the earth's atmosphere.

- Ceiling and Visibility- pilot can lose sight of runway.
- Precipitation- pits costly heat-absorbing tiles (labor intensive replacement process).
- Lightning (both natural and triggered)- can cause electrical problems and damage tiles as above (can also come from detached cirrus anvils < 3hrs old).
- Winds- shuttle could be blown off course.

This list of hazards due to thunderstorms highlights the need for forecasters to pay close attention to the development, dissipation, and movement of thunderstorms (Bauman and Businger, 1996).

2.2 Numerical Modeling

2.2.1 Introduction

With the advancement of computers over the last sixty years, the application of numerical methods to weather prediction has provided a valuable tool to forecasters. By representing the atmosphere with differential equations and then applying numerical methods to solve these equations, we have found that a reasonable approximation of the future state of the atmosphere can be attained. Current computer technology allows numerical weather forecasting on a global scale that at the beginning of the century was only a dream.

2.2.2 History

In 1904, Vilhelm Bjerknes suggested that given certain initial atmospheric variables it should be possible, in principle, to predict the future state of these variables (Haltiner and Williams, 1980). The first attempt to apply this theory was in 1922 by Lewis F. Richardson. Using finite difference forms of dynamic equations to represent fluid motions of the atmosphere, he attempted a 6-hr forecast. The forecast ended in complete failure, but Richardson knew why. The primary reasons for his forecast's failure were due to the lack of upper-air data and the poor quality of initial data at the surface. Besides data problems, the process was also tainted by the state of knowledge of numerical methods and dynamic meteorology of the time (Holton, 1992). Richardson foresaw that the complications due to observational and computational difficulties might one day be resolved, but he didn't foresee the development of super-computers which make numerical weather forecasting feasible (Fleagle and Businger, 1980). According to Richardson, "Perhaps someday in the dim future it will be possible to advance the computations faster than the weather advances and at a cost less than the saving to mankind due to the information gained. But that is a dream." (Richardson, 1922)

The first realization of this dream came in the late 1940's when a team of experts led by Jule Charney implemented a numerical weather forecast on the newly developed electronic computer. Using the advanced knowledge gained over the previous two decades, Charney noted some serious difficulties in using Richardson's equations. The main problem that Charney noted was that Richardson's equations allowed non-meteorologically significant waves to enter the solution. By simplifying the equations, Charney was able to produce a reasonable 24-hr forecast of the 500-mb height field

(Charney et al., 1950). Unfortunately, operational forecasts were still not possible, for it took nearly 24 hours to assimilate the data and output the forecast.

In the late 1950's, the first operational numerical forecasts based on a three-layer barotropic model were produced from a joint venture between the U.S. Air Force Air Weather Service, the U.S. Weather Bureau, and the Navy Weather Service (Ray, 1986). As time progressed, increasing knowledge of the atmosphere, better communication of meteorological data, new methods of performing numerical calculations, and better computers led to large advances in the field of numerical weather prediction. The early 1960's saw the first useful predictions of sea-level pressure, and the 1970's saw the birth of spectral transform methods, which are pervasive in global modeling. Since the 1970's, the field of numerical weather prediction has grown by leaps and bounds. Continued advances in computing power have led to highly complicated grid point and spectral models with high resolution and advanced physical parameterizations. The dreams of Bjerknes have obviously been far surpassed; we must wonder what advances lurk in the future to surpass our dreams.

2.2.3 Modeling Fundamentals

In order to describe a particular numerical forecast model as completely as possible, it is typical to classify the model based on three different categories. These categories highlight the basic concepts that set each model apart from the others and are based on the numerical method used to solve the equations, the scale at which the model is designed to provide information, and the type of equations used in the mathematical

model. Each of these categories is discussed below along with a discussion of initial and boundary conditions.

2.2.3.1 Numerical Method Used

The governing equations used to represent the atmosphere generally take the form of partial differential equations (PDE's) that need to be replaced by other functions in order to obtain solutions. The two main methods used are the finite difference method and the spectral method. The basics of each method, along with some considerations for each, are presented below.

In the finite difference approach, variables of interest are defined at points on a grid usually of equal spacing. Some models use a simple approach of defining each variable at every point, while others use a staggered grid in which variables are defined at alternating points. Staggering the grid offers some advantages when defining the finite difference formulas used in solving the PDE's (Arakawa and Lamb, 1977). The most common method to convert the continuous PDE's into algebraic equations is to represent the derivatives in the PDE's as truncated Taylor series. Depending on how the expansion is carried out, the dependent variable can be represented by either a centered, forward, or backward difference in either time or space. For example, in a backward time-centered space (BTCS) difference, the time derivative would be solved in terms of the past few values (how many depends on the order of the derivative), and the space derivative would depend on values to either side of the grid point of interest (again, the number of points to either side depends on the order of the derivative).

Finite difference methods can also be classified as either explicit or implicit. In an explicit finite difference, the dependent variable is defined in terms of known values (as in the BTCS case). Explicit schemes often suffer from numeric instability in which errors grow exponentially with continued calculations. Another possible problem with some explicit schemes is the introduction of a spurious mode, known as the computational mode, into the solution (Holton, 1992). Various methods to account for stability and computational mode problems have been devised, but are too lengthy to discuss here. The advantage of explicit schemes is that they are less computationally expensive than implicit schemes. When the dependent variable is defined in terms of other unknown variables the scheme is known as implicit. Since more than one variable is unknown, the equations must either be solved iteratively or through matrix methods. While implicit schemes are generally stable, they often suffer from errors in the phase and amplitude of the solution (Haltiner and Williams, 1980). As with explicit schemes, methods to get around some of the problems with implicit schemes have been devised. Other sources of errors inherent to finite difference methods relate to the truncating of the Taylor series, the step size of the time and distance steps used in performing the calculations, amplitude errors that occur when trying to represent waves smaller than about five times the grid spacing, and the inability to represent waves smaller than two times the grid spacing (Haltiner and Williams, 1980).

Spectral models use truncated orthogonal functions, called basis functions, to convert the PDE's into ordinary differential equations (ODE's). Common choices for basis functions are Fourier series, when the dynamic equations are in Cartesian coordinates, and spherical harmonics, when the dynamic equations are in spherical

coordinates (Holton, 1992). By using numerical methods to solve the ODE's, we are able to forecast the phase and amplitude of the basis functions (Holton, 1992). Since these functions now represent the phase and amplitude of the response, the values for the variables are now continuous, whereas in finite difference forms the values are only defined at grid points. One of the main benefits of using spectral models is that horizontal derivatives are evaluated exactly without the use of finite differences; this makes the calculation of horizontal advection highly accurate (Haltiner and Williams, 1980). Also, truncation errors in the space differencing schemes are eliminated by the use of spectral methods.

Since spectral methods use continuous functions, they are especially well suited for domains with periodic boundaries. This is why spectral models are generally always used in global modeling. Another reason spectral methods work so well in global modeling is that there is no need to map the spherical earth to a grid because all of the functions are already defined in spherical coordinates.

The method described above is widely used in spectral modeling; however, it has a few downfalls. This method works very well for linear equations and often outperforms finite difference methods of similar resolution (Haltiner and Williams, 1980). The problem lies in the fact that when this method is used to solve nonlinear equations, the number of simultaneous equations that need to be solved increases as the square of the number of modes (number of individual waves around the earth) that the model is attempting to resolve (due to the multiplication of series). This fact made spectral methods unmanageable for higher resolution models. To overcome this downfall, the 1970's saw the development of the spectral transform method. Through a

series of transformations, this method reduces the number of variables in the equations and therefore severely decreases the amount of terms that need to be multiplied together (Orszag, 1970). Because of the spectral transform method high-resolution spectral models are possible.

2.2.3.2 Scale of Interest

Currently, the scale of most models fall into one of three categories: global, synoptic, or mesoscale. Global models are generally spectral models that focus on forecasting large-scale features such as the long wave pattern, the jet stream, and upper air wind patterns. To get “the big picture” of weather systems like frontal positions, synoptic scale models are used. These models usually cover a domain roughly the size of North America and are best for analyzing features with a length scale of approximately 2000 kilometers (Ray, 1986). To properly model smaller phenomena like thunderstorms, mesoscale models are used. These models are generally high-resolution finite difference models that focus on phenomena with length scales less than 1000 kilometers. The size of the smallest phenomena that can be modeled depends on the grid resolution of the model. Due to the higher resolution and more complex physical parameterizations of mesoscale models, the domain size is usually limited by computer power.

2.2.3.3 Equations Used

The form of the equations used to represent the atmosphere determines the possible types of weather phenomena that can be modeled. Different ways to categorize

equation sets include barotropic models, baroclinic models, filtered models, and primitive equation models. Filtered models are usually based on simpler equation sets like baroclinic or barotropic models, but they are re-formulated to make up for some of the shortcomings of the unfiltered model. The more complicated primitive equation models permit the widest range of modeled phenomena but are the hardest to implement and require the most computer resources. Hydrostatic and non-hydrostatic versions of each model are also possible.

From the information presented above, it is now possible to provide an example using these naming conventions. The eta model, the basis for this thesis, is now more fully described as a non-hydrostatic mesoscale primitive equation model.

2.2.3.4 Initial and Boundary Conditions

In order for a numerical model to output a forecast, the initial meteorological variables within the model's domain must be specified (Haltiner and Williams). Data that is ingested from various sources (rawinsonde, NEXRAD, satellite, aircraft, etc.) are quality controlled and manipulated to follow certain constraints and provide a suitable representation of the initial state of the atmosphere. The sophistication of the initialization process usually determines the amount and type of data that go into this initialization. Newer methods allow both synoptic data (data taken simultaneously at selected locations on the hour) and asynoptic data (data taken between synoptic observations, at other locations, or both) to be considered in the initialization process.

The initialization process also interpolates the data from the point of observation to the grid points of the model.

There are two types of boundaries of concern to numerical weather prediction: physical and computational. At the surface of the earth, a physical boundary, mathematical difficulties exist due to the need for meteorological variables to terminate there within certain constraints. For example, wind speeds need to go to zero at the earth's surface. Computational boundaries exist at the edge of the model's domain. At these lateral boundaries, "the solution should depend continuously on the boundary conditions so that small changes in the boundary values produce only small changes in the solution (Haltiner and Williams, 1980)." Computational boundary problems exist for the lateral boundaries of finite difference models and windowed spectral models, but don't exist in global spectral models since they have periodic boundaries. As mentioned earlier, the method of differencing determines which points go into the calculation of the derivatives. If the model uses a forward difference in the interior of the domain, as the computations proceed toward the right edge there is a point (the number of grid points from the edge depends on the order of the Taylor series) at which it is no longer possible to continue calculations. Therefore, to perform calculations near a boundary a combination of differencing methods is usually necessary.

2.2.4 The Mesoscale Eta Model

The majority of the information given in this section on the eta model is from Black (1994). The mesoscale eta model is a numerical weather prediction model that derives its name from the high resolution of the model and the vertical coordinate system

that the model is based on. The eta vertical coordinate was first suggested by Mesinger in 1984 to reduce errors that occur in calculating the pressure gradient force, advection, and horizontal diffusion along steeply sloped coordinate surfaces. This coordinate system is defined by

$$\eta = \left(\frac{p - p_T}{p_{sfc} - p_T} \right) \left[\frac{p_{ref}(z_{sfc}) - p_T}{p_{ref}(0) - p_T} \right], \quad (1)$$

where p_T is the pressure of the top level in the domain (currently 25 mb in the mesoscale eta model), p_{sfc} and z_{sfc} are the pressure and height of the model's lower boundary, and p_{ref} is a reference pressure that is a function of height above sea level. Since the equation is set up to normalize the eta coordinate, the values for eta range from unity at the surface to zero at the top of the domain. The advantage of the relationship given by equation 1 is that eta surfaces remain relatively flat in areas of steeply sloped terrain.

NCEP currently produces four runs per day of the mesoscale eta model (00Z, 03Z, 12Z, and 18Z). Due to different methods of initialization, and various changes being implemented only to certain run times (usually the 03Z and 18Z runs are changed together and the 00Z and 12Z runs are changed together), only the 03Z information will be described since it is the focus of this research. However, most of the information given is common to all runs.

The 03Z eta model currently has a horizontal resolution of 32-km (see Figure 4) and 45 levels (see Figure 5) of varying thickness in the vertical (Rogers, 1998). This pairing was introduced into the operational suite of NCEP products on 3 June, 1998. Before June, the mesoscale eta model had 29-km resolution and 50 vertical layers. As a

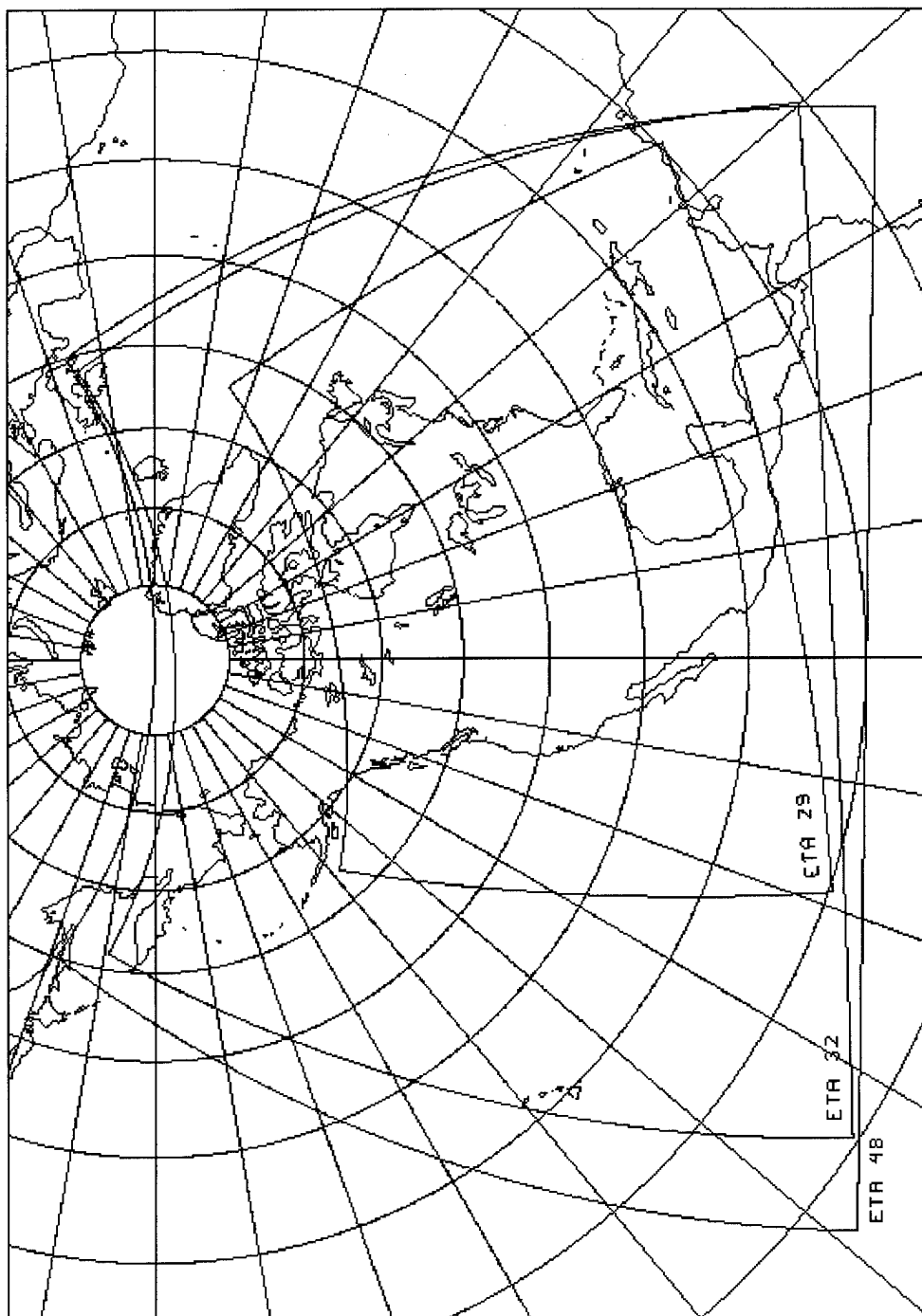


Figure 4 Map showing the 29-km, 32-km, and 48-km eta domains

Eta Model 45—Layer Distribution

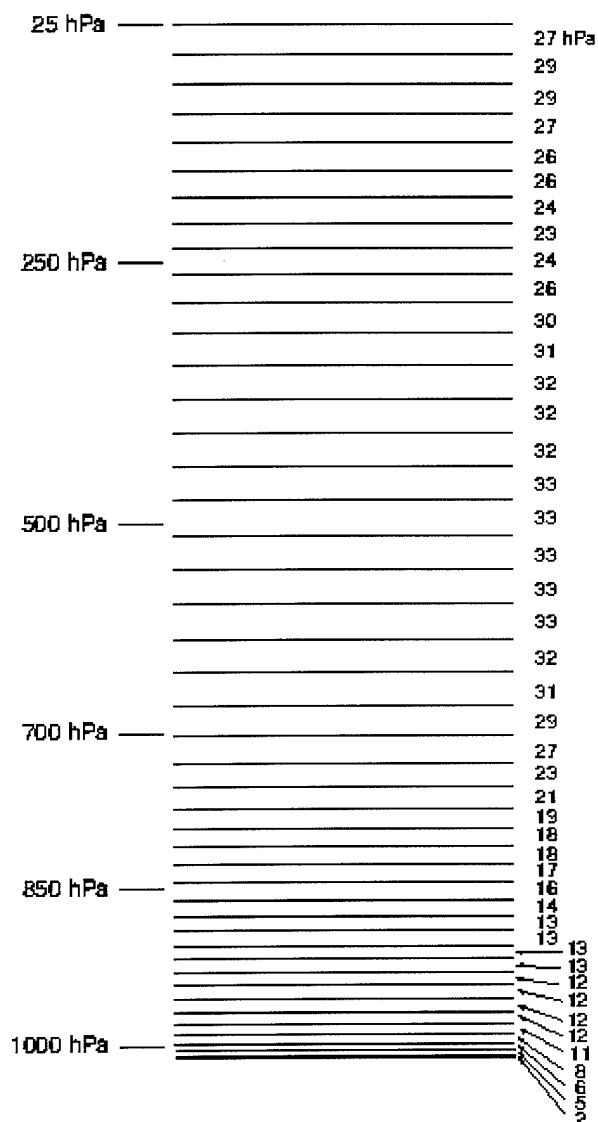


Figure 5 45 Vertical Levels of 32-km eta. Left margin shows approximate pressure levels corresponding to the layers of the eta model, and the right margin shows the approximate pressure difference between each eta level.

result of this change in resolution, the data for this thesis contains both versions of output. Relative maxima in the vertical resolution of the layers exist close to the surface and near the tropopause to allow better resolution in those areas.

As mentioned before, the mesoscale eta model is based on finite difference forms of the physical and dynamical atmospheric equations. The grid layout used in the model is referred to as a semi-staggered Arakawa E grid which is rotated onto a tangential Lambert conformal map to minimize the convergence of the meridians at the center of the domain. A sample of the Arakawa E grid is shown in Figure 6 below.

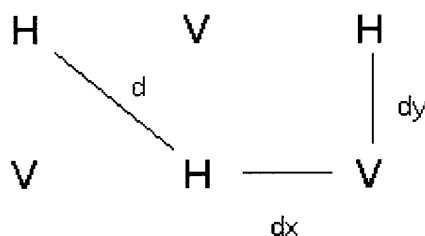


Figure 6 Semi-Staggered Arakawa E-grid

In figure 6, each H represents a mass point (such as temperature or moisture) and each V represents values of the horizontal wind. The distance $\sqrt{2d}$ is given as the resolution of the model (Arakawa and Lamb, 1977).

Topography is represented in the model using a silhouette step-mountain method. Through a series of averaging the elevations within a region, the actual height of the surface is raised or lowered to correspond to the closest eta layer interface. Because of

this, the height of the topography around CCAS usually takes on values between 0 feet and 20 feet above sea level.

The main prognostic variables of the eta model are temperature, u and v wind components, specific humidity, surface pressure, turbulent kinetic energy, cloud water, and cloud ice. In order to initialize the model, the Eta Data Assimilation System (EDAS) takes in data from the following sources:

- Rawinsondes
- Pibal winds
- Dropwindsondes
- Wind profilers
- Surface land temperature and moisture
- Oceanic surface data
- Aircraft winds
- Satellite cloud-drift winds
- Satellite precipitable water retrievals
- Aircraft temperature data
- Surface winds over land
- Velocity Azimuth Display (VAD) winds from NEXRAD
- Satellite oceanic surface winds

The 03Z run then uses the 00Z initialization panel as the first guess and begins to run a three-dimensional variational analysis (3-d VAR) on the data it has assimilated in the past three hours (from 00Z to 03Z). The object of using 3-d VAR is to map the input data to the 32-km eta domain with minimum analysis error by taking as many observations and past forecasts as possible into consideration (Rogers, 1998). Soil moisture is currently

initialized using climatological values because of the lack of standard moisture observations around the United States (Staudenmaier, 1996). Cloud water and cloud ice are also initialized differently than the other parameters and require a “spin-up” time before saturation can occur and precipitation can form. Since these parameterizations aren’t fully developed in the early forecast periods, care must be taken in analyzing the output of these fields.

To limit inertial gravity wave formation, a forward-backward scheme is used in the gravity wave adjustment phase. This technique eliminates the need for explicitly taking out waves with small wavelengths (Haltiner and Williams, 1980). The stability criterion for this scheme allows for relatively large time steps compared to other methods and also produces no computational mode. Vertical advection is accomplished by using a Euler-backward time scheme and a centered space scheme. The only exception to this is in the calculation of specific humidity where a piecewise linear method is used. Horizontal advection is accomplished through a combination of a modified Euler-backward scheme and a scheme developed by Janjic especially for horizontal advection on the E grid. For the advection of water vapor, a shape preserving scheme is used.

Convective processes are handled by a Betts-Miller-Janjic scheme (BMJ). This scheme gradually relaxes the temperature and moisture profiles toward a profile that has been observed in nature. The BMJ model is set up to handle both shallow and deep convective processes in order to model convective precipitation (Staudenmaier, 1996). Other schemes included in the model include a Mellor-Yamada scheme to model turbulent kinetic energy, a radiation scheme developed by the Geophysical Fluid Dynamics Laboratory (GFDL), visibility schemes, and cloud ice and water schemes.

In 1997, and again in 1998, the Applied Meteorology Unit (AMU) at CCAS was contracted to evaluate the usefulness of the mesoscale eta model to forecasting at the Cape. Both subjective and objective studies were performed on data from the 1996 and 1997 warm seasons. While these data sets fall into the timeframe that the 29-km version was being used, it is felt that due to the similarities in resolution and physics between the models, the results should apply equally well to the new 32-km version (Manobianco and Nutter Part II, 1998). The subjective portion of the study was performed to quantify the value added by the model forecasts to specific phenomena like thunderstorms, sea breezes, and fronts (NASA CR-205409, 1997). Results of the subjective verification suggest that the eta model is capable of forecasting organized convection but still lacks the resolution necessary to resolve individual air-mass type thunderstorms (Manobianco and Nutter Part II, 1998). Objective verification was performed on various individual parameters such as temperature, moisture, and winds at all levels. Results of this portion of the AMU study show that soundings generated by the model are generally too warm, too dry, and too stable, which also indicates that the model may not properly forecast thunderstorms. The AMU studies show that the model generally provides good forecast guidance, but until the resolution increases and more mesoscale information is incorporated in the model, the forecasting of individual thunderstorms is not possible (NASA CR-1998-207910).

2.2.5 Previous Work

Through the years, many different studies have been performed relating to thunderstorms around the Cape. These include many different sea breeze studies, flow

regime studies, and the development of the NPTI to name a few. A brief description of some of these works is presented.

2.2.5.1 The NPTI

One of the most important studies to the operational forecasting of thunderstorms at the Cape is that from which the NPTI was developed. Using multiple regression techniques on thirteen years of 12Z soundings from the Cape, Charles Neumann developed an index using variables found to be statistically important in the development of thunderstorms. He considered over 250 different variables from the soundings for inclusion in the model, of which only five survived to be included in the final model. Since non-linear trends in these variables were found to be important, Neumann included second and third order polynomials in these five variables to help explain as much of the variance as possible. The five variables in the equations are the 850-mb wind, the 500-mb wind, the mean relative humidity between 800 and 600-mb, the Showalter stability index (SSI), and the Julian day number. As a result of the regression analysis, it was determined that each month of the warm season had sufficiently different regression coefficients to warrant a separate equation. For a complete description of the formulation of the NPTI, see NOAA Technical Memorandum NWS SOS-8 (Neumann, 1971).

The probability of thunderstorms given by the NPTI has been shown to perform only slightly better than persistence (Kelly, 1998; Wohlwend, 1998; Howell, 1998). However, due to the lack of a significantly better objective forecasting tool the NPTI remains in operational use at the Cape. The NPTI is currently only used for forecasting

thunderstorms less than 24 hours out. Attempts to quantify its usefulness in forecasting beyond the present day (Wohlwend, 1998) by using data from the mesoscale eta model have resulted in inconclusive results. The NPTI was calculated from the model data for the days of this research. The index developed by this research will be compared to the NPTI and persistence as a measure of its value-added to the forecasting process.

2.2.5.2 Other Studies

Various studies in the past have been performed to assess the predictors that are important to thunderstorm development around the Cape. The two discussed here, sea breeze studies and flow regime studies, have been widely researched and prove to be important factors in thunderstorm development. The mesoscale eta model should be able to resolve the flow regime, but the resolution may not yet be high enough to model the smaller spatial scales of the sea breeze.

Recent sea breeze studies include those by Cetola (1997) and Zhong and Takle (1993). Both two-dimensional and three-dimensional sea breeze models were considered in these studies to determine the locations of the sea breeze front and areas of possible thunderstorm development. Both studies showed the sea breeze to be an important factor in the development of thunderstorms. Low-level moisture convergence, low-level vorticity, and low-level upward vertical motions were considered in this thesis as possible indicators of the presence of a sea breeze front.

Some recent flow regime studies were performed by Bauman and Businger (1996), Kelly (1998), Lopez and Holle (1987), and less recently by Neumann (1970). A

brief description of the importance of each of these works is presented below. Bauman and Businger proposed a higher thunderstorm probability when the low-level winds are from the southwest. The utility in rotating the coordinate system to parallel the coast of the Cape was demonstrated by Lopez and Holle. Kelly proposed the value of using the K-index as a thunderstorm predictor. Neumann's studies point to the 3000' winds as the best single thunderstorm predictor. Each of these findings were implemented into this thesis and analyzed as predictors in thunderstorm development.

3. Methodology

3.1 Overview

The aim of this research was the development of a model-based index capable of outperforming persistence and the NPTI. The first step was the collection of the model and observational data. After the model data were collected, variables were extracted from the data files and either used as-is, or used to develop other variables. The data set was then divided into many different model-building and validation sets in order to apply a logistic regression to each. The fit of each regression model was assessed and validation was performed on each of the validation data sets. The index developed through the logistic regression procedure was then compared to persistence and the NPTI using standard verification and skill techniques to assess its forecasting ability.

3.2 Data Processing

The observational and forecast data used in this research came from two different sources. Observational data for KTTS were received over the World Wide Web from the Air Force Combat Climatology Center (AFCCC) as a text file. A Fortran program was then written to find which days had thunderstorms on station (defined as thunder heard by the KTTS weather observer). The output of the program was a file with a "0" for days without a thunderstorm, and a "1" for days with a thunderstorm. As a quality control, about half of the values in the output file were checked to make sure they corresponded to the observations.

The forecast data consisted of 114 files (one file per day) of the 33 hr forecast from the 03Z run of the mesoscale eta model between the period of 1 May to 14 Sep, 1998. Twenty-one days of missing data were scattered over the four-and-a-half month period of the data set. Each of the data files came in a binary gridded format as described in Dey (1996) and contained all model output variables. Though the model runs at a resolution of 32-km, (29-km before June 3), NCEP stores the files on a 40-km Lambert conformal map projection. Software incorporated into the GEneral Meteorological PAcKage (GEMPAK) was used to extract the variables at the four nearest grid points surrounding KTTS (NW corner at 28.88N, -80.86W, SE corner at 28.55N, -80.40W). This step was also accomplished through the use of various UNIX scripts which automated the extraction procedure for multiple variables and millibar levels: the end result of this step was over 30,000 files.

Following Djuric (1994), the value of each variable at KTTS was found using a distance weighted average from KTTS to each of the four nearest grid points. The value at KTTS, B , is given by

$$B = \frac{\sum W_i Z_i}{\sum W_i}, \quad (2)$$

where the Z_i are the observations at each of the $i=4$ grid points and W_i is given by

$$W_i = \exp\left(-\frac{d_i^2}{c}\right). \quad (3)$$

The values of d_i are the distances from the grid points to KTTS, and c is defined by Djuric as a "suitably determined constant that controls the effective influence of an observation on the B value." Using the ratio of c to the average grid spacing of the

example given in Djuric's book, to the c for this grid and grid spacing (40-km) the value of c was estimated to be $26.5 (*10^3 \text{ m}^2)$. The final value of c was found by graphically generating ten different isopleth patterns, assigning values to the four grid points and interpolating the value from the isopleths nearest to KTTS manually, and determining what value of c produced the same value as the interpolated value at KTTS. The value of c that produced the most consistent results was $27.9 (*10^3 \text{ m}^2)$. The formulas above were then incorporated into a Fortran program that converts each of the 30,000 files (one file per day, per variable, per level) with the four grid points into one file per month, per variable, per level containing only the value at KTTS.

Averaging the wind direction had to be handled differently than the other variables. As an example, consider the following. If the two western most grid points had winds from 350 degrees, and the two eastern most points had winds from 030 degrees, then the averaging technique described above would result in a wind direction from 237 degrees. In reality, the wind direction should average from around 360 degrees. To account for this, if the grid had wind directions from both the northeast (NE) and northwest (NW) quadrants then 360 was added to the wind direction value in the NE quadrant before the averaging was accomplished. For the example given, the 030 direction would be converted to 390 in order to perform the averaging. If the averaging produced a value above 360, then 360 was subtracted from the average (putting KTTS wind direction in the NE quadrant); otherwise, the average was left as is (wind direction at KTTS in the NW quadrant). In the case of the example, the resultant direction was from 364 degrees, or 004 degrees after 360 is subtracted: a more realistic value for the directions given.

One thing to note before using the data retrieved from the model is that no attempts are made to account for model errors or biases. All of the forecast data is therefore taken to be correct as output from the NWP model. This is known as the perfect prognosis assumption (Wilks, 1995). If the forecast generated by the model really is "perfect," then we should expect that the index generated by the regression equations will also provide a good forecast.

Layer-averaged values used as predictors were calculated from a technique adapted from AWS/TR-83/001 standard programming guide (Duffield and Nastrom, 1983) which accounts for the logarithmic structure of the atmosphere. Relative humidity, specific humidity, cloud ice, and cloud water averages were computed and quality control checked by comparison to the arithmetic average. Summary statistics were also calculated to ensure the maximum and minimum values fell within acceptable limits.

The Showalter stability index (SSI) was also calculated for each day by use of a Fortran program using techniques from AWS/TR-83/001. Ten percent of the output SSI's were checked through graphical means on a Skew-T, Log-p diagram to ensure the values were correct. Values of SSI were used in computation of the NPTI and as a variable in the logistic regression.

As mentioned above, the forecast data is stored on a Lambert conformal map. This has repercussions when trying to decode the wind direction and speed at the grid points. Fortunately, GEMPAK incorporates software that automates the conversion of winds from the Lambert map to a north-relative coordinate system. As a quality control check, the wind rotation was accomplished by use of the equations that NCEP uses to encode the winds to the Lambert map. These values were quality checked against the

values generated by GEMPAK and against wind plots also generated by GEMPAK. The u and v wind components for the 850-mb and 500-mb levels were decomposed from the wind direction and speed for use in the NPTI calculation and as variables for the logistic regression.

Calculation of the NPTI was accomplished by writing a Mathcad program following the Fortran program of Neumann (1971). Input variables were the u and v wind components at the 850-mb and 500-mb levels, the SSI, the layer-averaged 800-mb to 600-mb relative humidity, and the Julian day number. Using data from Wohlwend (1998) the program was tested on three days of data. Values output from the program matched to Wohlwend's results to within 2%.

Various forms of persistence were tested for significance in the logistic regression model. A diagram of the persistence form that proved the most significant with respect to likelihood ratio scores, t-test values, and Wald statistics (see section 3.3.1 on logistic regression) is included as Figure 7. This form weighs the current and next-day thunderstorm persistence value more or less heavily depending on whether there was a thunderstorm on the previous day.

In order to decrease the size of the data files, variables that were represented in exponential notation were multiplied by an appropriate number so that they could be represented in decimal form. The variables that required this modification and the value that they were multiplied by are absolute vorticity ($\times 10^5$), omega ($\times 10^5$), cloud ice ($\times 10^6$), specific humidity ($\times 10^4$), and moisture convergence ($\times 10^8$). After the addition of the various persistence predictors, SSI, wind components, and many other variables, the final number of different variables in the matrix was 250.

The next step was to break up the data set by months and then perform a univariate logistic regression on each variable for each month. The Wald score and -2 Log L scores were retained from each regression in order to assess which variables were statistically significant to thunderstorm occurrence. As described in the logistic regression section below, these scores follow a chi-square distribution. Variables that produced p-values for the Wald scores and -2 Log L scores of .30 or less for at least three of the five months were retained into a separate file for further evaluation. Mickey and Greenland (1989) suggest the use of a higher statistical level than the traditional .05 as a screening criterion in the early stages of model-building. This higher level helps identify values that may interact with other variables to produce better statistical significance of the model. Of the 250 variables in the original matrix, only 94 produced the required p-value for retention in the reduced variable file set.

Each month of data was then separated into a model-building set and a validation set in order to develop a model for each month of data. Members in the validation set were chosen randomly by a computer program so that 80% of the data remained in the model-building set, and the remaining 20% went into the validation set. By building the model on one set of data and then validating it on another, it is possible to evaluate the predictive ability of the chosen model (Neter et al., 1989). One downfall to this is that the model-building set must be sufficiently large in order to provide a reliable model. A general guideline used by statisticians is to have the number of observations in the model-building set be 6-10 times more than the number of independent variables included in the model (Neter et al., 1989). These guidelines are based on the principle that a model will

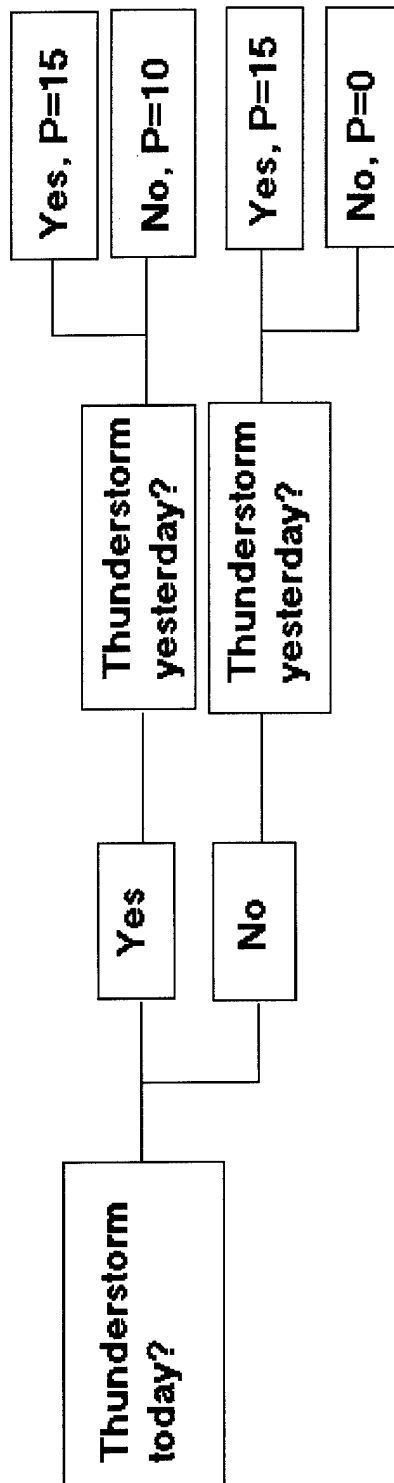


Figure 7 Two-day Persistence

be overfit when the number of predictors is large compared to the number of observations (Wilks, 1995). Using stepwise logistic regression on the 94 remaining variables, an attempt was made to determine which variables were important to thunderstorm formation for each month. Unfortunately, after taking out data for validation, the small sample size of each model-building set resulted in a statistically insignificant model (as indicated by the $-2 \text{ Log } L$ scores and the extremely high standard errors in the estimated coefficients).

Since breaking up the data by month resulted in the data sets being too small, the full data set was then divided into three periods. The dates of these divisions are 1 May- 15 Jun, 16 Jun-31 Jul, and 1 Aug- 14 Sep. Each trimester of data was labeled T1, T2, and T3 respectively. From each of these new sets, three sets of model-building and validation data sets were built using randomly chosen days from each set in a 80/20 ratio as before. The three model-building and validation sets of each trimester were also added together into three model-building sets and validation sets of the full data set. Also, the different sets of T1 and T2, as well as T2 and T3, were added together to form additional model-building and verification sets (named T1T2 and T2T3 respectively). Hence, the resulting model and validation data sets were labeled T1-A, T1-B, T1-C, T1-All, T2-A through T2-All (as in T1), T3-A through T3-All, T1T2-A through T1T2-All, T2T3-A through T2T3-All, and Full-A through Full-All. Each "A", "B" and "C" suffix for the validation sets denotes a different 20% of data removed from the larger set from which it came (T1T2 for instance). The "ALL" suffix denotes the full period of data with no data removed for validation (all of the data for T2T3 for instance).

A stepwise logistic regression was then performed on each of these eighteen model-building sets, to determine which variables were significant, and physically meaningful, to the occurrence of a thunderstorm. The importance of the variables that came out of the stepwise regression was based on the Wald chi-square, the odds ratio, the t-value, and the residual deviance (as described in section 3.3.1 on logistic regression).

The beta coefficients output by the statistical packages were then applied to the data in the validation sets using the logistic regression formula. Forecast thunderstorm probability values output by the regression equations were then verified using standard verification and skill techniques calculated using a 2 X 2 contingency table. The verification techniques used are discussed below in the section on verification.

3.3 Statistical Methods

A perfect forecast may never be achieved due to dynamical chaos inherent to the atmosphere (Wilks, 1995). How close we are able to come to a perfect forecast depends on how much we can reduce the uncertainty that the weather problem presents us. Statistical methods provide a basis to explain at least part of this uncertainty. Statistical procedures used in this thesis include logistic regression analysis for building a statistical model, and verification statistics to test the performance of the model.

3.3.1 Logistic Regression

Logistic regression has its roots in epidemiology where it is widely used to model the binary outcome variable of whether a person has a disease or not. The utility in using this method in other fields, such as meteorology, is beginning to gain wider acceptance.

In this thesis, logistic regression techniques have been used to model the probability of thunderstorm occurrence since the presence or absence of a thunderstorm can be represented using a dichotomous variable.

The form of logistic regression used in this thesis is based on equation 1.1 from Hosmer and Lemshow (1989). For multiple logistic regression the equation is

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}, \quad (4)$$

where $\pi(x)$ is known as the logit transformation, the x_n are the observations, and the β_n are the logistic regression coefficients. The forecast value is then determined by the equation $y(x) = \pi(x) + \varepsilon$, where ε is the error with a mean of zero, and a variance of $\varepsilon = \pi(x)[1 - \pi(x)]$. The major benefit of using logistic regression versus linear regression is that equation 2 restricts the forecast values to between 0 and 1 where linear regression has no such bounds. Another major difference between logistic and linear regression is that the errors follow a binomial distribution versus a normal distribution. As a result, the least squares method and residuals usually employed in linear regression modeling are no longer applicable.

Since the least squares method isn't applicable, the logistic regression model uses what is known as the maximum likelihood function. "This function expresses the probability of the observed data as a function of the unknown parameters" (Hosmer and Lemshow, 1989). The maximum likelihood function determines the maximum likelihood estimator which models each estimate to agree most closely with the observed data. The likelihood estimators are then incorporated into another function resulting in what is known as the deviance (D). Deviance is found by the equation:

$$D = -2 \sum_{i=1}^n \left[y_i \ln \left(\frac{\hat{\pi}_i}{y_i} \right) + (1 + y_i) \ln \left(\frac{1 - \hat{\pi}_i}{1 - y_i} \right) \right] \quad (5)$$

Where the $\hat{\pi}_i$ are the maximum likelihood estimates and the y_i are the i^{th} forecast values.

Deviance is the logistic regression equivalent of residuals in linear regression.

In order to test the significance of the addition of independent variables to the model, the difference in the deviance of the model without the variables and with the variables is calculated. Since the deviance is calculated using the likelihood function, the model's significance can also be found using the formula $G = -2 \ln(L)$, where L is the ratio of the likelihood of the model without the variables to the likelihood of the model with the variables, and G is known as the likelihood ratio test (also known as $-2 \text{ Log } L$). An important property of the likelihood ratio test is that the values of G follow a chi-square distribution with $n-1$ degrees of freedom where n is the number of parameters in the model. Since the likelihood ratio test follows this distribution, it is easy to calculate p-values for G . By choosing a level of significance of 95%, we are able to judge the significance of having included the variables in the model. If the p-value is 0.05 or less, then the addition of the variables to the model is judged to be significant.

The method above describes how to assess the significance of the overall model. Once the overall model is judged to be significant, an assessment of the individual variables in the model can be made. One method to judge the significance of the estimated regression coefficients is to calculate a statistic called the Wald test (Hosmer and Lemshow, 1989). Values of the Wald test follow a chi-square distribution with $n+1$ degrees of freedom and therefore can be interpreted using a p-value in the same manner

as in the likelihood ratio test. Another statistic to use in determining the significance of the predictors is the odds ratio. Simplification of the odds ratio formula results in the equation $\psi = \exp(\beta_1)$, where ψ is the odds ratio and β_1 is the beta-coefficient of the independent variable. This statistic has found wide use because it approximates how much more likely a predictor is to be important on days that have a thunderstorm ($x=1$) than among days with no thunderstorm ($x=0$). A value of one indicates an equal likelihood between $x=0$ and $x=1$. For a value of $\psi = 1.32$, the variable is said to be 32% more likely to be present when $x=1$ than when $x=0$.

3.3.2 Verification

The process of determining the quality of a forecast is known as forecast verification (Wilks, 1995). This process involves matching a set of forecasts to the observations to which they pertain in order to determine the effectiveness of the forecast method. Historically, verification of forecasts has been rather controversial due to the numerous methods of verification that are available (Panofsky and Brier, 1968). Because of this, a number of different verification techniques are used in order to give a broad view of the relation between the observations and forecasts. The verification methods used in this thesis are presented below.

A 2 X 2 contingency table approach was used to verify the ability of the logistic regression model to forecast the occurrence of thunderstorms. This method requires categorical data as input to the contingency table matrix before calculations can be made. Since the output values of the logistic regression model fall in the range of zero to one, a cutoff value is needed so that the forecast values can be converted to zeros and ones

(Neter et al., 1989). The forecast data were visually compared to the observations to see what cutoff might work best. From the distribution of forecasts to observations, it was determined that a cutoff anywhere between 0.4 and 0.6 would produce the same overall result. So, 0.5 was chosen as the cutoff value; values 0.5 or greater were converted to a one, and values less than .5 were converted to a zero. The contingency table is set-up as shown in Figure 8.

A Mathcad template was written that converted the forecast data to categorical data and then entered the data into a contingency table. Before verification statistics can be said to be significant, the dependence between the observations and forecasts in the contingency table must be determined (Wilks, 1995). Dependence indicates that the contingency table was created by related versus random events. The assessment of dependence was determined through the use of Yates' continuity corrected chi-square statistic, and Fisher's exact test (Everitt, 1992).

Since the chi-square distribution is based on a continuous probability distribution, and the contingency table uses discrete events, it is suggested to use the following formulation of the chi-square distribution when dealing with small sample sizes (Everitt, 1992)

$$X_c^2 = \frac{N(|ad - bc| - .5N)^2}{(a+b)(c+d)(a+c)(b+d)} \quad (7)$$

This is known as Yates' continuity corrected chi-square. It follows the standard chi-square distribution with one degree of freedom; therefore, a value of 3.84 or higher (or $P(\text{chi-square}) \leq 0.05$) indicates dependence between the observations and forecasts.

		Observed		
		Yes	No	
Forecast	Yes	a	b	a+b
	No	c	d	c+d
		a+c	b+d	n = a + b + c + d

Figure 8 2X2 Contingency Table- “a” represents an event that was both forecast and observed, “b” represents an event that was forecast but not observed. Boxes “c” and “d” follow similarly. Sums on the right and bottom margin are the marginal totals.

Fisher's exact test is used when expected cell counts in the contingency table are less than about 5 (as most were with the sample sizes used in this thesis). To find the value of Fisher's exact test, the following formula is used to compute all possible combinations of the contingency table that produce the same marginal totals (column totals and row totals)

$$P = \frac{(a+b)! (a+c)! (c+d)! (b+d)!}{a! b! c! d! N!} \quad (8)$$

Each combination of cells that produces the same marginal totals has a corresponding value of P associated with it. The value of Fisher's exact test is the sum of the P's from each combination that produces the same marginal totals. The value output by Fisher's exact test is a p-value, so a value less than 0.05 is assumed significant at at least a 95% significance level.

With dependence established, it is now possible to calculate accuracy and skill scores. The first of these indicates the fraction of correct forecasts and is given by the hit rate (*HR*)

$$HR = \frac{a+d}{n} \quad (9)$$

A perfect forecast results in a hit rate of one and lower values indicate worse forecasts. Threat score yes (no) indicates the number of correct "yes" ("no") forecasts divided by the total number of times the event was forecast and/or observed. Threat score yes (*TSY*) and threat score no (*TSN*) are given by

$$TSY = \frac{a}{a+b+c} \quad \text{and} \quad TSN = \frac{d}{d+b+c} \quad (10/11)$$

A perfect forecast results in threat score values of one while lower values indicate worse forecasts. The probability of detection (POD) is the percent of time that the forecast event occurred and was also forecast ($PODY$ for “yes” forecasts, and $PODN$ for “no” forecasts). These are given by

$$PODY = \frac{a}{a+c} \quad \text{and} \quad PODN = \frac{d}{d+b} \quad (12/13)$$

The POD of a perfect forecast is one while lower values indicate worse forecasts. The proportion of forecast events that fail to occur is given by the false alarm rate. False alarm rate yes ($FARY$), and false alarm rate no ($FARN$) are given by

$$FARY = \frac{b}{a+b} \quad \text{and} \quad FARN = \frac{c}{c+d} \quad (14/15)$$

A low false alarm rate is desirable; therefore, false alarm rates of zero are produced by a perfect forecast.

To see if an event was overforecast or underforecast, the bias is computed. Bias is simply the ratio of “yes” forecasts to “yes” observations; it is given by

$$BIAS = \frac{a+b}{a+c} \quad (16)$$

A bias of one indicates that the event was forecast as often as it was observed. A bias greater (less) than one means that the event was overforecast (underforecast). Since the bias doesn’t take the correspondence of observations and forecasts into account, it isn’t considered an accuracy measure.

Forecast skill is a measure of the relative accuracy between a set of forecasts and some set of reference forecasts. Common choices for the accuracy measures (A) are those mentioned above. The reference forecast (A_{ref}) used in this thesis was the 24-hour

persistence. A perfect forecast (A_{perf}) is with respect to the accuracy measure used. The formula for skill score using a reference forecast is

$$SS = \frac{A - A_{ref}}{A_{perf} - A_{ref}} 100. \quad (17)$$

A skill score of zero indicates no improvement over the reference forecast, and a score of 100 indicates a perfect forecast. Skill scores less than zero indicate that the forecast method being evaluated performs worse than the reference forecast. Two other skill scores commonly used are the Kuipers skill score (KSS) and the Heidke skill score (HSS). They are computed using the formulas

$$KSS = \frac{(ad - bc)}{(a + c)(b + d)} \quad (18)$$

and,

$$HSS = \frac{2(ad - bc)}{(a + c)(c + d) + (a + b)(b + d)}. \quad (19)$$

Perfect scores for the Heidke and Kuipers scores receive a score of one, forecasts equivalent to a random forecast receive a score of zero, and forecasts worse than that achieved from a random forecast receive negative scores (Wilks, 1995).

The Brier score (BS) and Brier skill score (BSS) were also computed. The Brier score is computed by averaging the squared differences between the forecast probabilities (not the categorical forecasts as above) and the subsequent binary observations (Wilks, 1995). Values for the Brier score range from zero, for a perfect forecast, to one with higher values indicating less accurate forecasts. The Brier score is given by

$$BS = \frac{1}{N} \sum_{k=1}^N (y_k - o_k)^2 100. \quad (20)$$

Where the y_k are the forecasts, and the o_k are the observations. The Brier skill score is a result of using the Brier score in the skill score formula (equation 16), and it is interpreted in the same manner as the other skill scores.

4. Analysis and Results

4.1 Overview

This chapter presents the findings of this research. The variables that were found to be important to the occurrence of thunderstorms are presented along with an explanation of their physical importance. The indices developed (one for each of the data sets) using these variables is referred to as the Eta Thunderstorm Index (ETI). Since a contingency table approach is used to verify the results, the dependence of the observations on the forecasts is determined using Yates' corrected chi-square and Fisher's exact test. These statistics are discussed to show the significance of the contingency tables and the significance of any comparison between persistence, the ETI, and the NPTI. Last, a comparison of the verification statistics calculated for each of the forecasts is presented along with an explanation of the significance of the forecasts.

4.2 Variables Found to be Important

As mentioned before, logistic regression was used to find the variables that were important to thunderstorm development. Unfortunately, the statistics calculated for T1, T2, and T3 proved to be statistically insignificant, so no further analysis was performed on these sets. The set labeled T1 had only four days of thunderstorms in its 36 days of data, which was insufficient for building a reliable model. Sets T2 and T3 produced insignificant verification results on the validation data sets most likely due to the small sample size of the data sets (only 8 values in each set). The three remaining sets, T1T2,

T2T3, and the full set, produced the variables in Table 1 as important to thunderstorm occurrence.

Table 1 Variables Found to be Important to Thunderstorm Occurrence

T1T2	T2T3	FULL
2-Day Persistence	2-Day Persistence	2-Day Persistence
900mb Wind Direction (dir900)	850mb Wind Speed (spd850)	900mb Wind Direction (dir900)
RH at Freezing Level (fzrel)	500mb Wind V-Component Squared- $(v500)^2$	600mb Relative Humidity (rh600)
250mb Absolute Vorticity Squared- $(av250)^2$	575mb Specific Humidity (sh575)	250mb Absolute Vorticity (av250)
	325mb Cloud Ice (ci325)	Convective Available Potential Energy (CAPE)
	950mb Specific Humidity (sh950)	

The variables in Table 1 were each run through the logistic regression procedure for the three validation subsets of each of the three periods (T1T2, T2T3, and Full). The estimated beta coefficients for each period and the coefficients for the full set of data (model and validation sets added together) are presented in Table 2, Table 3, and Table 4.

Table 2 Beta Coefficients for T1T2

	Intercept	Persistence	dir900	fzrel	$(av250)^2$
T1T2-A	-15.12289	.43427	.0159	.13232	.01192
T1T2-B	-13.08726	.36272	.01859	.09775	.00961
T1T2-C	-10.1404	.29357	.01062	.08287	.00234
T1T2- ALL	-12.34987	.33948	.014069	.09796	.01052

*Intercept- intercept from logistic regression, Persistence- 2-day persistence indicator, dir900- 900-mb wind direction, fzrel- freezing level relative humidity, $(av250)^2$ - square of the 250-mb absolute vorticity.

Table 3 Beta Coefficients for T2T3

	Intercept	Persist	spd850	(v500) ²	sh575	ci325	sh950
T2T3-A	-34.211	.61892	-1.1418	.21296	.3343	-.06738	.15203
T2T3-B	-28.108	.522	-.81636	.25095	.36386	-.06395	.09435
T2T3-C	-17.577	.35773	-.3383	.13805	.20857	-.06697	.05342
T2T3-ALL	-18.173	.36287	-.41732	.11579	.17611	-.04684	.07035

*Intercept- intercept from logistic regression, Persist- 2-day persistence indicator, spd850- 850-mb wind speed, (v500)²- square of the 500-mb wind v-component, shXXX- specific humidity at the XXX-mb level, ci325- 325-mb cloud ice content.

Table 4 Beta Coefficients for Full Set

	Intercept	Persistence	dir900	rh600	av250	CAPE
FULL-A	-11.66527	.3413	.00969	.09406	.08887	.00049
FULL-B	-12.67858	.35045	.01197	.10373	.06671	.0006
FULL-C	-9.80365	.31437	.00596	.08375	.0118	.00035
FULL-ALL	-11.80008	.329446	.009203	.093957	.102774	.000617

*Intercept- intercept from logistic regression, Persistence- 2-day persistence indicator, dir900- 900-mb wind direction, rh600- relative humidity at 600-mb, av250- 250-mb absolute vorticity, CAPE- convective available potential energy.

With the tables of important variables and beta coefficients, it is now possible to interpret the physical relation between each variable and thunderstorms. The sign of the beta coefficient indicates the relation between the variable and the occurrence of a thunderstorm. A positive (negative) beta coefficient means that higher (lower) values of the variable correspond statistically to thunderstorm occurrence. One important point common to all of the tables is that the range of beta coefficients for each predictor is rather small. This is desirable in that it indicates that the predictors for each data set have relatively the same effect. Tables of the residual deviance and t-value for each variable are included in Appendix B.

The two-day persistence is common to the three subsets of data as expected. This variable shows the statistical dependence of thunderstorm occurrence on whether

thunderstorms were present the previous days. When the same synoptic regime dominates an area for a period of time, as is common in the summer, we expect that similar weather conditions should appear in a number of successive days. Therefore, under these conditions it is expected (statistically) that days with no thunderstorms will be followed by a day without a thunderstorm, and days with thunderstorms will be followed by another day with a thunderstorm. This is quite evident in the data used for this thesis; after July first, thunderstorms come in average runs of 4.4, and days without thunderstorms come in average runs of 4.1.

Numerous studies have shown the 900-mb (or ~3000') wind direction to be one of the best indicators of thunderstorms near Cape Canaveral. From Neumann (1970), "the estimated probability of thunderstorm occurrence displays marked variation with 3000-foot wind direction. The minimum value of 10 percent with northwesterly winds contrasts rather sharply with the maximum of nearly 60 percent with winds from the southwest." Cetola (1997) and Lopez et al. (1984) verified this for more recent data and indicated higher preference for thunderstorms with winds from the southwest through west. With a relatively low occurrence of northwesterly winds (see Howell (1998) for wind rose information), the positive correlation of wind direction to thunderstorm occurrence indicates southwesterly to westerly winds to be of prime importance. The most likely explanation as to why this variable is important is that it relates to the orientation of the coastline at the Cape, and therefore affects the convergence patterns of the wind near the shuttle landing facility.

Positive correlation in the specific humidity at 575-mb, 950-mb, the relative humidity at 600mb, and the freezing level relative humidity are indicative of the depth of

the moist layer. Deep moist layers have been shown to be positively correlated with shower activity in the Florida peninsula (Frank and Smith, 1968). The presence of a deep moist layer could be indicative of upward moisture transport and therefore upward vertical motion which is a necessary factor to thunderstorm development. A question that begs to be asked is why the layer average values that were computed aren't in the list of variables in place of the individual levels. This occurs because of correlations in the data. Layer average values were found to be important, but when they were combined with the other model variables much of the deviance that they explained was already explained by other variables. The moisture variables used had less of a correlation with the other variables in the model and therefore made a larger contribution to explaining the deviance.

The 850-mb wind speed and the square of the v -wind component at 500mb have also previously been shown to relate to thunderstorm occurrence. The beta coefficient for the 850-mb wind speed shows a negative correlation with thunderstorms. Neumann (1971) also found this to be true. In terms of wind components, the speed (or magnitude) of the wind is given by $Speed = \sqrt{u^2 + v^2}$. The constants in Neumann's research indicate negative correlation in the square of both the u and v component at 850-mb, and therefore the speed as well. Relatively light winds and low vertical wind shear are two of the defining characteristics of single cell thunderstorms (Ray, 1986); the results of the regression support this theory.

A positive correlation in the square of the 500-mb v -wind component was found to be an important variable in both this thesis and in Neumann's work. This result "indicates that the effect of wind speed is not predominantly linear and direct but rather

parabolic" (Lopez et al., 1984). On 95% of the days with thunderstorms, the square of the 500-mb wind component was over 10 ms^{-1} , compared to only 30% of the days without thunderstorms having speeds over 10 ms^{-1} . Since the v -wind component relates to the meridional orientation, the square of this variable indicates an importance in the departure of the wind direction from the east and west. This result implies that the synoptic regime also plays an important role in thunderstorm occurrence.

The next variable found to be important, absolute vorticity, is the sum of relative vorticity and planetary vorticity. Since relative vorticity is usually small compared to planetary vorticity (for mid-latitude synoptic systems) we would expect that the values of absolute vorticity are going to be mostly positive (Holton, 1992). Lower values of absolute vorticity should relate to negative relative vorticity, and higher values should relate to positive vorticity. Therefore, the positive correlation indicates that positive vorticity at 250-mb is important to thunderstorm occurrence. This indicates that the magnitude of the counter-clockwise rotation in the upper atmosphere relates to the strength of upward vertical motions in the atmosphere. Stronger values of absolute vorticity could also relate to the position of the sub-tropical jet stream and its associated convergence/divergence zones.

The density of ice within the model layer at 325-mb was found to be negatively correlated to the occurrence of thunderstorms. Since the model data are forecasting the atmosphere at 12Z, this variable indicates that the presence of clouds at 325-mb hinders the development of thunderstorms throughout the day. The most likely explanation for why this variable is important is that the high clouds reduce the amount of surface heating thereby taking away a main trigger mechanism for thunderstorm development.

Convective available potential energy (CAPE) was also found to be positively correlated to thunderstorm occurrence. This variable is a measure of the positive buoyant energy available in the atmosphere to lift parcels. CAPE was found to only have a minimal impact on reducing the deviance for the full set of data but it was included since it did slightly reduce the deviance. Other stability indices also performed poorly, this is possibly a result of the eta model being too dry, warm, and stable (NASA CR-1998-207910, 1998).

The reason that the variables that are important to thunderstorm occurrence change from period to period isn't clear. Each period contains the two-day persistence, a variable relating to the winds, and a variable relating to the relative humidity as the strongest predictors. With the exception of the SSI and day number, these results are close to what Neumann found in his research. The importance of stability is weakly indicated in the full set by the CAPE variable, and Neumann himself indicated that day number was a weak predictor only included to maintain consistency from month to month (Neumann, 1971). A possible explanation as to why the variables differ is that this research didn't have a large enough sample size (more years of data) to resolve features important for the whole summer. Or, the important variables may actually change from month to month. Another possible explanation is that the eta model may inconsistently model the atmospheric variables resulting in vacillation of the important variables. It is also possible that other non-linear interactions in the data that weren't explored could prove to be more important. Without further research using more data, it is difficult to say why there is difference in the predictors.

4.3 Significance of Contingency Tables

Using the results of Yates' corrected chi-square and Fisher's exact test, the significance of the contingency tables for 24-hr persistence, the validation sets of the ETI, and the NPTI were assessed. The 24-hour persistence forecasts were proved to be related to the observations at a significance level of at least 92% for the full set, but proved insignificant for each validation set of T1T2 and T2T3 (T1T2-A through C, and T2T3- A through C).

Validation sets of the ETI were shown to have good overall significance. All but T1T2-C produced significance levels of 95% or greater. Set T1T2-C gives a significance level of 94.7% when Yates' continuity corrected chi-squared and Fisher's exact test are averaged. Since both tests measure the significance of the contingency table, and since both tests are considered conservative, we can conclude that T1T2-C is also significant at the 95% significance level.

None of the contingency tables produced by the NPTI data were found significant at the 95% significance level. Thus the NPTI proved to be a rather poor method of forecasting thunderstorms with little correlation between the forecasts and the observations. It should be noted that this is the NPTI forecast made using the eta model and not the NPTI from the atmospheric sounding that is in operational use at Cape Canaveral.

4.4 Verification Statistics

The verification statistics below are based on the contingency table approach mentioned earlier. Results are given only for cases in which Yates' corrected chi-square

and Fisher's exact test indicated that the contingency table was statistically significant. If persistence was found not to be statistically significant, then the skill score computed against the ETI was also insignificant and was not analyzed. The most significant results came from the three subsets of the full data set. Results from each data set are analyzed and discussed below.

The contingency tables produced using the data from T1T2-A, T1T2-B, and T1T2-C showed only the ETI to be statistically significant at the 95% significance level. The average hit rate was a gracious 85.4%. This, coupled with a 6% *FARN* and a relatively low Brier score, leads to the conclusion that this method is quite robust. One slight detriment exists in the value of the bias; the average value of 133% indicates a slight over forecasting for this data set. The other accuracy and skill computations also reflect positively on the forecasting ability of the model used for each subset of the data.

In subset T2T3-A, persistence and the ETI were found to be significant. However, both contingency tables had the same cell counts, so both methods produced the same statistics (so skill score = 0). In subset T2T3-B, none of the forecast methods proved to be statistically significant. Comparing the accuracy and skill measures of T2T3-A and T2T3-C, the average hit rate is 81.3%, the average *FARN* is 17%, and the average Brier score is a low 0.125. Average bias is close to 100% and therefore indicates neither over nor under forecasting of these models.

For all of the subsets of the full model, both persistence and the ETI were found to be significant. Because of this, we can compare the two by use of the skill score. The positive values of the skill score indicate the ETI does a better job of forecasting than persistence. One similarity worth mentioning is that the value of *PODN* for persistence

and the ETI is the same within each subset. An analysis of these contingency tables revealed that the "b" and "d" cells were the same in each table, which indicates that the "observed no" column in the contingency table is identical for both persistence and the ETI. Since the other skill and accuracy scores are higher for the ETI, the logical conclusion is that the ETI's formulation results in better forecasting of the "observed yes" column. Again, the average hit rate of 84.7%, the average Brier score of .113, and the relatively low bias of 116% suggest this method has merit in thunderstorm forecasting. Probably the most operationally significant forecast is the one that doesn't forecast a thunderstorm and a thunderstorm occurs. The statistics of this combination are represented in the probability of detection no accuracy measure (*PODN*). The average *PODN* for the ETI is only 8.6% compared to 17.8% for persistence. So, *PODN* is another indicator of the ETI's utility over persistence.

5. Summary and Recommendations

5.1 Overview

The previous chapters introduced, developed, and analyzed an index built of the mesoscale eta model for the purpose of thunderstorm forecasting. Even with a somewhat limited data set, the index showed skill in predicting thunderstorms at Cape Canaveral. Continued improvements in the mesoscale eta model will require a continued evaluation of predictors used in building the index. A larger database of mesoscale eta data should be analyzed in order to refine the index and possibly develop a dynamic index that changes from month to month. A summary of the findings of this research and some recommendations for further work are discussed below.

5.2 Summary

Mesoscale eta model data and observational data were collected for the warm season of 1998 and entered into a matrix for analysis. The data were then divided into three subsets demarcated by date: 1 May-15 Jun, 16 Jun-31 Jul, and 1 Aug-14 Sep. A stepwise logistic regression was performed on the full set of data and the three subsets to find which variables were important to the occurrence of thunderstorms at Cape Canaveral. Each of the data sets were then further divided into a model-building and validation set in an 80/20 ratio. The model-building sets were used to find the regression coefficients, which were then applied to the validation data sets.

The probabilities output by applying the regression equations to the validation data sets were then categorized as either no thunderstorms (if the probability was less

than 0.5) or thunderstorms (if the probability was greater than or equal to 0.5). The categorized data were then entered into two-by-two contingency tables for statistical analysis and for comparison to the NPTI and persistence. Analysis of the three subsets of data showed only the ETI to be consistently statistically significant at the 95% significance level. Skill and accuracy scores for the subsets of data indicate a relatively high amount of forecast ability for the ETI. The full data set proved both persistence and the ETI to be statistically significant at the 95% significance level. In each of the validation sets, the ETI had increased skill over persistence. The NPTI was shown to be insignificant at the 95% level for all data sets.

Though the verification results were favorable and the validation sets showed no signs of an overfit model, the small sample size make the results of this research only tentative. The results of this thesis are therefore of limited operational use, but they do provide a basis for further work.

5.3 Recommendations

The scourge of this thesis was definitely the data. Sources for archived eta data are rare, and when one is found the data sets are usually incomplete or contain periods of missing data. The amount of data received for this thesis was sufficient for only a preliminary study on the feasibility of using model data to build an index for forecasting thunderstorms. For a more definitive study on this topic, it is suggested that many more years of warm season data be included in the process. The representativeness of observational data could also be improved upon. This study only included observation data from the shuttle landing facility. Other sources such as radar data, satellite data, and

lightning data could be included to ensure a more representative pool of thunderstorm days. A greater pool of observational data and model data would help to smooth out the annual variability of the results and reduce the dependence of the results on the data set.

The persnickety nature of the NPTI as a result of its linear regression basis was surpassed by the logistic regression formulation of the ETI. Logistic regression proved to be a valuable tool for this type of application and should continue to be used for studies of this type. One possible improvement on the methods of this thesis is to include more interaction terms in the regression. The multiplication of variables by one another or the inclusion of cubic terms may result in a better model. Also, if more data can be obtained a regime-based index may prove worth looking into.

Another limiting factor to the success of this technique is the mesoscale eta model itself. With constant changes to the eta model, a study involving more than one warm season will result in having to use versions with differing resolutions, physical parameterizations, and initialization procedures. It is entirely possible that the variables that the model models as being important to thunderstorm occurrence change from version to version. Therefore, as the mesoscale eta model is updated, it will be necessary to continually adjust the regression coefficients to keep the forecast ability of the regression equations tuned. An extension of the methods used in this thesis to other forecast models such as the medium range forecast (MRF) might prove useful for forecasting past day-2. Continued effort on this topic is encouraged.

Appendix A: Variables Considered

This appendix contains the list of variables (and levels) tested for significance to thunderstorm occurrence in this thesis.

Absolute Vorticity (s^{-1})- 1000, 850, 750, 700, 500, 400, 250-mb's

Cloud Ice (kg/m^2)- 600 through 200 mb's in increments of 25 mb's

Cloud Water (kg/kg)- 1000 through 600-mb's in increments of 25 mb's

Dewpoint (K)- 1000 through 500-mb's in increments of 25 mb's, and 250-mb's

Temperature (K)- 1000 through 500-mb's in increments of 25 mb's, and 250-mb's

Omega (Pa/s)- 1000 through 500-mb's in increments of 25 mb's, and 250-mb's

Relative Humidity (%) - 1000 through 500-mb's in increments of 25 mb's, 400 mb's and 250-mb's

Layer Average Relative Humidity (%) - 300-500, 500-700, 600-800, 700-925, 800-1000 mb's

Specific Humidity (kg/kg)- Same levels as Relative Humidity

Layer Average Specific Humidity (kg/kg)- Same levels as layer avg. Relative Humidity

Cloud Water Layer Average (kg/kg)- 600-800, 700-900, 600-900, 800-1000 mb's

Cloud Ice Layer Averages (kg/m^2)- 200-400, 300-500, 400-600-mb's

Heights (gpm)- 1000, 925, 850, 700, 600, 500, 400, 250-mb's

Thickness (gpm)- 1000-500, 925-600, 700-400, 850-500, 500-250-mb's

Wind Direction (degrees)- 1000-700 mb's in 50 mb increments, 600, 500, and 250-mb's

Wind Speed (Knots)- Same levels as Wind Direction

U and V Wind Components (knots)- 850 and 500-mb's

Binary Directional Wind Shear- 850-500, 950-750 mb's

Directional Wind Shear- 850-500, 950-750 mb's

Wind Speed Shear- 700-500, 950-800-mb's

Squared Wind Speeds- 900, 850, 700, 500, 250-mb's

Convective Precipitation (kg/m^2) Non-Convective Precipitation (kg/m^2)

Total Precipitation (kg/m^2) CAPE (J/kg)

CINS (J/kg) Mean Sea Level Pressure (Pa)

K-Index (Celsius degrees) SWEAT Index (non-dimensional)

SSI (Celsius degrees) 10-Meter AGL Speed (knots)

10 Meter AGL Direction (degrees) Lifted Index (Celsius Degrees)

Helicity (m^2/s^2) Moisture Convergence (kg/kg/s)

Climatology Persistence

Day Number Categorical Winds

Categorical Persistence Freezing Level RH (%) and Height (gpm)

NPTI Categorical 10 Meter Wind Direction

Categorical 950, 900, and 850-mb Wind Direction

Appendix B: Percent Deviance Explained and T-Value Tables

Table B-1. T1T2-A: Percent Deviance Explained and T-Values

Variables	Persistence	dir900	fzrel	(av250) ²	Total Dev. Exp.
% Dev. Exp.	39.7	9.0	10.7	5.0	64.4
t-value	3.24	2.10	2.31	1.64	

*Persistence- 2-day persistence indicator, dir900- 900-mb wind direction, fzrel- freezing level relative humidity, (av250)²- square of the 250-mb absolute vorticity.

Table B-2. T1T2-B: Percent Deviance Explained and T-Values

Variables	Persistence	dir900	fzrel	(av250) ²	Total Dev. Exp.
% Dev. Exp.	43.2	12.2	7.6	2.4	65.4
t-value	3.38	2.08	2.19	1.17	

*Variables same as Table B-1.

Table B-3. T1T2-C: Percent Deviance Explained and T-Values

Variables	Persistence	dir900	fzrel	(av250) ²	Total Dev. Exp.
% Dev. Exp.	37.8	6.9	10.2	.1	55.0
t-value	3.30	1.34	2.21	.27	

*Variables same as Table B-1.

Table B-4. T2T3-A: Percent Deviance Explained and T-Values

Variables	Persist	spd850	v500 ²	sh575	ci325	sh950	Total Dev. Exp.
% Dev. Exp.	41.3	2.5	11.0	8.2	4.8	8.5	76.3
t-value	2.76	-2.40	2.42	2.48	-1.56	2.11	

*Persist- 2-day persistence indicator, spd850- 850-mb wind speed, (v500)²- square of the 500-mb wind v-component, shXXX-specific humidity at the XXX-mb level, ci325- 325-mb cloud ice content.

Table B-5. T2T3-B: Percent Deviance Explained and T-Values

Variables	Persist	spd850	v500 ²	sh575	ci325	sh950	Total Dev. Exp.
% Dev. Exp.	49.4	2.5	9.5	9.3	4.6	3.2	78.5
t-value	2.61	-2.12	2.08	2.22	-1.47	1.47	

*Variables same as Table B-4.

Table B-6. T2T3-C: Percent Deviance Explained and T-Values

Variables	Persist	spd850	v500 ²	sh575	ci325	sh950	Total Dev. Exp.
% Dev. Exp.	41.3	1.2	7.0	5.4	5.7	1.7	62.3
t-value	3.05	-1.66	1.94	2.24	-1.74	1.19	

*Variables same as Table B-4.

Table B-7. Full-A: Percent Deviance Explained and T-Values

Variables	Persist	dir900	rh600	av250	cape	Total Dev. Exp.
% Dev. Exp.	44.4	2.64	6.8	.7	.8	44.6
t-value	4.63	2.08	2.53	1.03	.96	

*Persistence- 2-day persistence indicator, dir900- 900-mb wind direction, rh600- relative humidity at 600-mb, av250- 250-mb absolute vorticity, CAPE- convective available potential energy.

Table B-8. Full-B: Percent Deviance Explained and T-Values

Variables	Persist	dir900	rh600	av250	cape	Total Dev. Exp.
% Dev. Exp.	50.1	4.2	7.0	.50	.75	62.55
t-value	4.39	2.13	2.30	.67	.91	

*variables same as Table B-7.

Table B-9. Full-C: Percent Deviance Explained and T-Values

Variables	Persist	dir900	rh600	av250	cape	Total Dev. Exp.
% Dev. Exp.	44.6	1.8	7.5	.01	.3	54.23
t-value	5.19	2.12	2.86	1.27	1.28	

*variables same as Table B-7.

Appendix C: Verification Statistics

Table C-1. Verification Statistics for T1T2-A

T1T2-A	Perfect Scores	Persistence	ETI		NPTI	
			Forecast	Skill	Forecast	Skill
HR %	100	75	87.5	50	43.8	-125
TSY %	100	33.3	66.7	50	10	-35
TSN %	100	71.4	83.3	41.7	40	-110
PODY %	100	50	100	100	25	-50
PODN %	100	83.3	83.3	0	50	-200
FARY %	0	50	33.3	33.3	85.7	-71.4
FARN %	0	16.7	0	100	33.3	-100
BIAS %	100	100	150		175	
HSS	1	.33	.71		-.20	
KSS	1	.33	.83		-.25	
BS	0	.25	.12		.23	
BSS %	100		53.1		8.6	

Table C-2. Verification Statistics for T1T2-B

T1T2-B	Perfect Scores	Persistence	ETI		NPTI	
			Forecast	Skill	Forecast	Skill
HR %	100	75	87.5	50	50	-100
TSY %	100	33.3	60	40	11.1	-33.3
TSN %	100	71.4	84.6	46.2	46.7	-86.7
PODY %	100	66.7	100	100	33.3	-100
PODN %	100	76.9	84.6	33.3	53.8	-100
FARY %	0	60	40	33.3	85.7	-42.9
FARN %	0	9.1	0	100	22.2	-144
BIAS %	100	167	167		233	
HSS	1	.35	.67		-.09	
KSS	1	.44	.85		-.23	
BS	0	.25	.13		.22	
BSS %	100		48.4		14.0	

Table C-3. Verification Statistics for T1T2-C

T1T2-C	Perfect Scores	Persistence	ETI		NPTI	
			Forecast	Skill	Forecast	Skill
HR %	100	68.8	81.3	40	56.3	-40
TSY %	100	28.6	57.1	40	30	2
TSN %	100	64.3	75	30	46.2	-50.8
PODY %	100	33.3	66.7	50	50	25
PODN %	100	90	90	0	60	-300
FARY %	0	33.3	20	40	57.1	-71.4
FARN %	0	30.8	18.2	40.9	33.3	-8.33
BIAS %	100	50	83.3		117	
HSS	1	.26	.59		.09	
KSS	1	.23	.57		.10	
BS	0	.31	.11		.23	
BSS %	100		65.4		27.1	

Table C-4. Verification Statistics for T2T3-A

T2T3-A	Perfect Scores	Persistence	ETI		NPTI	
			Forecast	Skill	Forecast	Skill
HR %	100	81.3	81.3	0	43.8	-200
TSY %	100	66.7	66.7	0	30.8	-108
TSN %	100	70	70	0	25	-150
PODY %	100	85.7	85.7	0	57.1	-200
PODN %	100	77.8	77.8	0	33.3	-200
FARY %	0	25	25	0	60	-140
FARN %	0	12.5	12.5	0	50	-300
BIAS %	100	114	114		143	
HSS	1	.63	.63		-.09	
KSS	1	.64	.64		-.10	
BS	0	.19	.14		.28	
BSS %	100		23.0		-49.2	

Table C-5. Verification Statistics for T2T3-B

T2T3-B	Perfect Scores	Persistence	ETI		NPTI	
			Forecast	Skill	Forecast	Skill
HR %	100	68.8	75	20	50	-60
TSY %	100	44.4	50	10	38.5	-10.8
TSN %	100	58.3	66.7	20	27.3	-74.5
PODY %	100	80	80	0	100	100
PODN %	100	63.6	72.7	25	27.3	-100
FARY %	0	50	42.9	14.3	61.5	-23.1
FARN %	0	12.5	11.1	11.1	0	100
BIAS %	100	160	140		260	
HSS	1	.38	.48		.19	
KSS	1	.44	.53		.27	
BS	0	.50	.23		.22	
BSS %	100		26.5		30.2	

Table C-6. Verification Statistics for T2T3-C

T2T3-C	Perfect Scores	Persistence	ETI		NPTI	
			Forecast	Skill	Forecast	Skill
HR %	100	68.8	81.3	40	62.5	-20
TSY %	100	44.4	66.7	40	53.8	16.9
TSN %	100	58.3	70	28	33.3	-60
PODY %	100	50	75	50	87.5	75
PODN %	100	87.5	87.5	0	37.5	-400
FARY %	0	20	14.3	28.6	41.7	-108
FARN %	0	36.4	22.2	38.9	25	31.3
BIAS %	100	62.5	87.5		150	
HSS	1	.38	.63		.25	
KSS	1	.38	.63		.25	
BS	0	.313	.11		.23	
BSS %	100		65.4		28.1	

Table C-7. Verification Statistics for Full-A

FULL-A	Perfect Scores	Persistence	ETI		NPTI	
			Forecast	Skill	Forecast	Skill
HR %	100	83.3	91.7	50	50	-200
TSY %	100	60	80	50	25	-87.5
TSN %	100	77.8	87.5	43.7	40	-170
PODY %	100	75	100	100	50	-100
PODN %	100	87.5	87.5	0	50	-300
FARY %	0	25	20	20	66.7	-167
FARN %	0	12.5	0	100	33.3	-167
BIAS %	100	100	125		150	
HSS	1	.63	.82		0	
KSS	1	.63	.88		0	
BS	0	.167	.08		.23	
BSS %	100		54.1		-40	

Table C-8. Verification Statistics for Full-B

FULL-B	Perfect Scores	Persistence	ETI		NPTI	
			Forecast	Skill	Forecast	Skill
HR %	100	75	79.2	16.7	54.2	-83.3
TSY %	100	45.5	54.5	16.7	31.3	-26
TSN %	100	68.4	72.2	12	42.1	-83.3
PODY %	100	71.4	85.7	50	71.4	0
PODN %	100	76.5	76.5	0	47.1	-125
FARY %	0	44.4	40	10	64.3	-44.6
FARN %	0	13.3	7.1	46.4	20	-50
BIAS %	100	129	143		200	
HSS	1	.44	.55		.14	
KSS	1	.48	.62		.19	
BS	0	.25	.15		.22	
BSS %	100		40.5		11.6	

Table C-9. Verification Statistics for Full-C

FULL-C	Perfect Scores	Persistence	ETI		NPTI	
			Forecast	Skill	Forecast	Skill
HR %	100	75	83.3	33.3	62.5	-50
TSY %	100	45.5	63.6	33.3	43.8	-3.1
TSN %	100	68.4	76.5	25.5	47.1	-67.6
PODY %	100	50	70	40	70	40
PODN %	100	92.9	92.9	0	57.1	-500
FARY %	0	16.7	12.5	25	46.2	-177
FARN %	0	27.8	18.8	32.5	27.3	1.82
BIAS %	100	60	80		130	
HSS	1	.46	.65		.26	
KSS	1	.43	.63		.27	
BS	0	.25	.11		.22	
BSS %	100		56.7		13.9	

Appendix D: Mathcad Template for Calculating NPTI

The following is a Mathcad template for calculating the NPTI for days in June

vrbs := READPRN("d:\mctemp\npti_vrbs.txt") file with NPTI variables

mo_con := READPRN("d:\mctemp\npti_mo_con.txt") file with monthly coefficients

C := READPRN("d:\mctemp\npti_consts.txt") file with NPTI coefficients

u v s t

vrbs format: daynum rh u500 v500 u850 v850 ssi rows(vrbs) = 114

const format: may jun jly aug sep (as in Neumann 71) cols(vrbs) = 8

0 1 2 3 4 i := 24.. 46

j := 0.. 113

w := 1 So this template only calculates NPTI for June. Need to change certain coefficients depending on the month of interest.

vrbs_{j,3} := vrbs_{j,3} · 1.9438445 vrbs_{j,5} := vrbs_{j,5} · 1.9438445 convert winds to knots

vrbs_{j,4} := vrbs_{j,4} · 1.9438445 vrbs_{j,6} := vrbs_{j,6} · 1.9438445

u₁ := vrbs_{1,3} s₁ := vrbs_{1,5} rh₁ := vrbs_{1,2} ssi₁ := vrbs_{1,7}

v₁ := vrbs_{1,4} t₁ := vrbs_{1,6} day₁ := vrbs_{1,1}

$$\begin{aligned} \text{fx1}_1 := & C_{0,w} + C_{1,w} \cdot s_1 + C_{2,w} \cdot t_1 + C_{3,w} \cdot s_1 \cdot t_1 + C_{4,w} \cdot (s_1)^2 + C_{5,w} \cdot (t_1)^2 + C_{6,w} \cdot (s_1)^3 \dots \\ & + C_{7,w} \cdot (s_1)^2 \cdot t_1 + C_{8,w} \cdot s_1 \cdot (t_1)^2 + C_{9,w} \cdot (t_1)^3 \end{aligned}$$

$$\begin{aligned} \text{fx2}_1 := & C_{10,w} + C_{11,w} \cdot u_1 + C_{12,w} \cdot v_1 + C_{13,w} \cdot u_1 \cdot v_1 + C_{14,w} \cdot (u_1)^2 + C_{15,w} \cdot (v_1)^2 + C_{16,w} \cdot (u_1)^3 \dots \\ & + C_{17,w} \cdot (u_1)^2 \cdot v_1 + C_{18,w} \cdot u_1 \cdot (v_1)^2 + C_{19,w} \cdot (v_1)^3 \end{aligned}$$

$$fx3_1 := C_{20,w} + C_{21,w} \cdot rh_1 + C_{22,w} \cdot (rh_1)^2 + C_{23,w} \cdot (rh_1)^3$$

$$fx4_1 := C_{24,w} + C_{25,w} \cdot ssi_1 + C_{26,w} \cdot (ssi_1)^2$$

$$fx5_1 := C_{27,w} + C_{28,w} \cdot day_1 + C_{29,w} \cdot (day_1)^2$$

The following programs fix the results in the case of strong easterly winds

$$fx1_1 := \begin{cases} err \leftarrow t_1 + .3 \cdot s_1 + 15.4 \\ .01 & \text{if } err \leq 0 \\ fx1_1 & \text{otherwise} \end{cases}$$

$$fx2_1 := \begin{cases} err \leftarrow v_1 + 5 \cdot u_1 + 78.9 \\ .01 & \text{if } err \leq 0 \\ fx2_1 & \text{otherwise} \end{cases}$$

The following programs make adjustments to the output due to the problems with using linear regression methods (REEP) as discussed in the thesis text.

$$fx1_1 := \begin{cases} fx1_1 & \text{if } 0 \leq fx1_1 < 1 \\ .01 & \text{if } fx1_1 < 0.01 \\ .99 & \text{if } fx1_1 \geq 1 \end{cases}$$

$$fx2_1 := \begin{cases} fx2_1 & \text{if } 0 \leq fx2_1 < 1 \\ .01 & \text{if } fx2_1 < 0.01 \\ .99 & \text{if } fx2_1 \geq 1 \end{cases}$$

$$fx3_1 := \begin{cases} .01 & \text{if } (rh_1 - 15) < 0 \\ fx3_1 & \text{otherwise} \end{cases}$$

$$fx4_1 := \begin{cases} .01 & \text{if } fx4_1 < 0.01 \\ fx4_1 & \text{otherwise} \end{cases}$$

Calculate the NPTI

$$P_Jun := mo_con_{0,w} + mo_con_{1,w} \cdot fx1 + mo_con_{2,w} \cdot fx2 + mo_con_{3,w} \cdot fx3 \dots \\ + mo_con_{4,w} \cdot fx4 + mo_con_{5,w} \cdot fx5$$

$$P_Jun_1 := \begin{cases} .01 & \text{if } P_Jun_1 < 0 \\ P_Jun_1 & \text{otherwise} \end{cases}$$

sample of output: multiplied by 100 to get percent

tt := 35.. 45

Julian day #	NPTI
$\text{vrbs}_{\text{tt},1}$	$\text{P_Jun}_{\text{tt}} \cdot 100$
166	4.687
167	35.312
169	42.38
171	38.291
172	23.827
174	52.479
176	53.98
177	52.07
178	43.81
179	39.023
180	22.694

Appendix E: Mathcad program for 99% Ellipse Winds

This program determines if the winds are in the 99% ellipse of the data set for the NPTI

vrbs := READPRN ("d:\mctemp\npti_vrbs.txt") Read in NPTI variables

i := 0.. 113 index

u_i := vrbs_{i,3} · 1.9438445 s_i := vrbs_{i,5} · 1.9438445 redefine variables for easy id and
change winds to knots

v_i := vrbs_{i,4} · 1.9438445 t_i := vrbs_{i,6} · 1.9438445

day_i := vrbs_{i,1}

ma := 0.. 23 may

ju := 24.. 46 jun

jl := 47.. 74 jly

au := 75.. 101 aug

se := 102.. 113 sep

row numbers for the days of
the month in my data set

$$a5 := \begin{bmatrix} 45.06 \\ 33.46 \\ 28.04 \\ 30.72 \\ 40.06 \end{bmatrix} \quad b5 := \begin{bmatrix} 34.26 \\ 25.64 \\ 23.18 \\ 22.89 \\ 27.15 \end{bmatrix} \quad xh5 := \begin{bmatrix} 12.336 \\ 5.065 \\ 2.048 \\ .995 \\ 2.560 \end{bmatrix}$$

$$yk5 := \begin{bmatrix} -1.539 \\ .511 \\ 1.7 \\ 2.202 \\ .268 \end{bmatrix} \quad ct5 := \begin{bmatrix} .82511 \\ .98741 \\ .97630 \\ .83962 \\ .83772 \end{bmatrix} \quad st5 := \begin{bmatrix} .56497 \\ .15816 \\ .21644 \\ .54317 \\ .46020 \end{bmatrix}$$

$$ct8 := \begin{bmatrix} .84897 \\ .89180 \\ .98686 \\ .84989 \\ .83772 \end{bmatrix} \quad st8 := \begin{bmatrix} .52844 \\ .45243 \\ .1616 \\ .52696 \\ .54610 \end{bmatrix} \quad a8 := \begin{bmatrix} 33.79 \\ 30.58 \\ 25.88 \\ 26.77 \\ 35.78 \end{bmatrix}$$

$$b8 := \begin{bmatrix} 24.06 \\ 21.78 \\ 18.58 \\ 19.26 \\ 24.93 \end{bmatrix} \quad xh8 := \begin{bmatrix} -.12 \\ 1.902 \\ 2.146 \\ .061 \\ -2.573 \end{bmatrix} \quad yk8 := \begin{bmatrix} .596 \\ 3.257 \\ 4.941 \\ 4.359 \\ 1.887 \end{bmatrix}$$

$$xp5_{ma} := (u_{ma} - xh5_0) \cdot ct5_0 + (v_{ma} - yk5_0) \cdot st5_0 \quad \text{may}$$

$$yp5_{ma} := (v_{ma} - yk5_0) \cdot ct5_0 - (u_{ma} - xh5_0) \cdot st5_0$$

$$sum5_{ma} := \frac{(xp5_{ma})^2}{(a5_0)^2 + \frac{(yp5_{ma})^2}{(b5_0)^2}}$$

$$xp5_{ju} := (u_{ju} - xh5_0) \cdot ct5_0 + (v_{ju} - yk5_0) \cdot st5_0 \quad \text{jun}$$

$$yp5_{ju} := (v_{ju} - yk5_0) \cdot ct5_0 - (u_{ju} - xh5_0) \cdot st5_0$$

$$sum5_{ju} := \frac{(xp5_{ju})^2}{(a5_0)^2 + \frac{(yp5_{ju})^2}{(b5_0)^2}}$$

$$xp5_{jl} := (u_{jl} - xh5_0) \cdot ct5_0 + (v_{jl} - yk5_0) \cdot st5_0 \quad \text{jul}$$

$$yp5_{jl} := (v_{jl} - yk5_0) \cdot ct5_0 - (u_{jl} - xh5_0) \cdot st5_0$$

$$sum5_{jl} := \frac{(xp5_{jl})^2}{(a5_0)^2 + \frac{(yp5_{jl})^2}{(b5_0)^2}}$$

$$xp5_{au} := (u_{au} - xh5_0) \cdot ct5_0 + (v_{au} - yk5_0) \cdot st5_0 \quad \text{aug}$$

$$yp5_{au} := (v_{au} - yk5_0) \cdot ct5_0 - (u_{au} - xh5_0) \cdot st5_0$$

$$sum5_{au} := \frac{(xp5_{au})^2}{(a5_0)^2 + \frac{(yp5_{au})^2}{(b5_0)^2}}$$

$$xp5_{se} := (u_{se} - xh5_0) \cdot ct5_0 + (v_{se} - yk5_0) \cdot st5_0 \quad \text{sep}$$

$$yp5_{se} := (v_{se} - yk5_0) \cdot ct5_0 - (u_{se} - xh5_0) \cdot st5_0$$

$$sum5_{se} := \frac{(xp5_{se})^2}{(a5_0)^2 + \frac{(yp5_{se})^2}{(b5_0)^2}}$$

winds_99_pct_{i,0} := day_i

500 mb winds

$$\text{winds_99_pct}_{\text{ma},1} := \begin{cases} 0 & \text{if } (\text{sum5}_{\text{ma}} - 1) < 0 \\ 5 & \text{otherwise} \end{cases} \quad \sum_{\text{ma}=0}^{23} \text{winds_99_pct}_{\text{ma},1} = 10$$

so there were two days in
may with winds out of the
99% ellipse

$$\text{winds_99_pct}_{\text{ju},1} := \begin{cases} 0 & \text{if } (\text{sum5}_{\text{ju}} - 1) < 0 \\ 5 & \text{otherwise} \end{cases} \quad \sum_{\text{ju}=24}^{46} \text{winds_99_pct}_{\text{ju},1} = 0$$

$$\text{winds_99_pct}_{\text{jl},1} := \begin{cases} 0 & \text{if } (\text{sum5}_{\text{jl}} - 1) < 0 \\ 5 & \text{otherwise} \end{cases} \quad \sum_{\text{jl}=47}^{74} \text{winds_99_pct}_{\text{jl},1} = 0$$

$$\text{winds_99_pct}_{\text{au},1} := \begin{cases} 0 & \text{if } (\text{sum5}_{\text{au}} - 1) < 0 \\ 5 & \text{otherwise} \end{cases} \quad \sum_{\text{au}=75}^{101} \text{winds_99_pct}_{\text{au},1} = 0$$

$$\text{winds_99_pct}_{\text{se},1} := \begin{cases} 0 & \text{if } (\text{sum5}_{\text{se}} - 1) < 0 \\ 5 & \text{otherwise} \end{cases} \quad \sum_{\text{se}=102}^{113} \text{winds_99_pct}_{\text{se},1} = 0$$

850-mb winds

$$xp8_{ma} := (s_{ma} - xh8_0) \cdot ct8_0 + (t_{ma} - yk8_0) \cdot st8_0 \quad \text{may}$$

$$yp8_{ma} := (t_{ma} - yk8_0) \cdot ct8_0 - (s_{ma} - xh8_0) \cdot st8_0$$

$$sum8_{ma} := \frac{(xp8_{ma})^2}{(a8_0)^2 + \frac{(yp8_{ma})^2}{(b8_0)^2}}$$

$$xp8_{ju} := (s_{ju} - xh8_0) \cdot ct8_0 + (t_{ju} - yk8_0) \cdot st8_0 \quad \text{jun}$$

$$yp8_{ju} := (t_{ju} - yk8_0) \cdot ct8_0 - (s_{ju} - xh8_0) \cdot st8_0$$

$$sum8_{ju} := \frac{(xp8_{ju})^2}{(a8_0)^2 + \frac{(yp8_{ju})^2}{(b8_0)^2}}$$

$$xp8_{jl} := (s_{jl} - xh8_0) \cdot ct8_0 + (t_{jl} - yk8_0) \cdot st8_0 \quad \text{jul}$$

$$yp8_{jl} := (t_{jl} - yk8_0) \cdot ct8_0 - (s_{jl} - xh8_0) \cdot st8_0$$

$$sum8_{jl} := \frac{(xp8_{jl})^2}{(a8_0)^2 + \frac{(yp8_{jl})^2}{(b8_0)^2}}$$

$$xp8_{au} := (s_{au} - xh8_0) \cdot ct8_0 + (t_{au} - yk8_0) \cdot st8_0 \quad \text{aug}$$

$$yp8_{au} := (t_{au} - yk8_0) \cdot ct8_0 - (s_{au} - xh8_0) \cdot st8_0$$

$$sum8_{au} := \frac{(xp8_{au})^2}{(a8_0)^2 + \frac{(yp8_{au})^2}{(b8_0)^2}}$$

$$xp8_{se} := (s_{se} - xh8_0) \cdot ct8_0 + (t_{se} - yk8_0) \cdot st8_0 \quad \text{sep}$$

$$yp8_{se} := (t_{se} - yk8_0) \cdot ct8_0 - (s_{se} - xh8_0) \cdot st8_0$$

$$sum8_{se} := \frac{(xp8_{se})^2}{(a8_0)^2 + \frac{(yp8_{se})^2}{(b8_0)^2}}$$

$$\text{winds_99_pct}_{\text{ma},2} := \begin{cases} 0 & \text{if } (\text{sum8}_{\text{ma}} - 1) < 0 \\ 8 & \text{otherwise} \end{cases} \quad \sum_{\text{ma} = 0}^{23} \text{winds_99_pct}_{\text{ma},2} = 8$$

one day with 850-mb
winds out of the ellipse

$$\text{winds_99_pct}_{\text{ju},2} := \begin{cases} 0 & \text{if } (\text{sum8}_{\text{ju}} - 1) < 0 \\ 8 & \text{otherwise} \end{cases} \quad \sum_{\text{ju} = 24}^{46} \text{winds_99_pct}_{\text{ju},2} = 0$$

$$\text{winds_99_pct}_{\text{jl},2} := \begin{cases} 0 & \text{if } (\text{sum8}_{\text{jl}} - 1) < 0 \\ 8 & \text{otherwise} \end{cases} \quad \sum_{\text{jl} = 47}^{74} \text{winds_99_pct}_{\text{jl},2} = 0$$

$$\text{winds_99_pct}_{\text{au},2} := \begin{cases} 0 & \text{if } (\text{sum8}_{\text{au}} - 1) < 0 \\ 8 & \text{otherwise} \end{cases} \quad \sum_{\text{au} = 75}^{101} \text{winds_99_pct}_{\text{au},2} = 0$$

$$\text{winds_99_pct}_{\text{se},2} := \begin{cases} 0 & \text{if } (\text{sum8}_{\text{se}} - 1) < 0 \\ 8 & \text{otherwise} \end{cases} \quad \sum_{\text{se} = 102}^{113} \text{winds_99_pct}_{\text{se},2} = 0$$

winds_99_pct =

	0	1	2
1	122	0	0
2	123	0	0
3	124	0	0
4	125	0	0
5	126	0	0
6	127	0	0
7	128	0	0
8	129	5	8
9	130	5	0
10	131	0	0
11	132	0	0
12	133	0	0
13	134	0	0
14	135	0	0
15	136	0	0

So, the 500 and 850-mb winds were out of the 99% ellipse (defined by Neumann's data) for day 129 and the 500-mb wind was out for day 130.

Bibliography

- AWS/TR-79/006: *The use of the Skew T, Log P Diagram in Analysis and Forecasting*. Air Weather Service, Scott AFB, IL, Revised Mar. 1990.
- Arakawa, A, and V. Lamb, 1977: Computational Design of the Basic Dynamical Processes of the UCLA General Circulation Model. *Methods in Computational Physics*, Vol 17, Academic Press, 174-265, 337 pp.
- Bauman, William H., and S. Businger, 1996: Nowcasting Convective Activity for Space Shuttle Landings at Kennedy Space Center, Florida. *Bull Amer. Met. Soc.*, **10**, 2295-2303.
- Black, Thomas L. 1994: The New NMC Mesoscale Eta Model: Description and Forecast Examples. *Wea. Forecasting*, **9**, 265-278.
- Bluestein, H. B., 1993: *Synoptic Dynamic Meteorology in Midlatitudes, Volume II*. Oxford University Press, Inc., 283 pp.
- Cetola, J. D., 1997: *A Climatology of the Sea Breeze at Cape Canaveral, Florida*. Masters Thesis, Florida State University, 56 pp.
- Charney, J. G, R. Fjortoft, J. von Neumann, 1950: *Numerical Integration of the Barotropic Vorticity Equation*. *Tellus*. **2**, 237-254.
- Dey, Clifford H., 1996: NCEP Office Note 388, GRIB (edition 1): *The WMO Format for the Storage of Weather Product Information and the Exchange of Weather Product Messages in Gridded Binary Form*. NCEP Central Operations 100 pp.
- Djuric, Dusan, 1994: *Weather Analysis*. Prentice-Hall Inc., p 38.
- Duffield, George F., and G. D. Nastrom, 1983: AWS/TR-83/001, *Equations and Algorithms for Meteorological Applications in Air Weather Service*. Air Weather Service, 58 pp.
- Everitt, B.S., 1992: *The Analysis of Contingency Tables*. 2nd ed., Chapman and Hall, 164 pp.
- Fleagle, Robert, and Joost Businger, 1980: *An Introduction to Atmospheric Physics*. 2nd ed, Academic Press, Inc., 151.
- Frank, N. L., and D. L. Smith, 1968: On the Correlation of Radar Echoes over Florida with Various Meteorological Parameters. *J. Appl. Meteor.*, **7**, 712-714.

- Haltiner, G. J., and R. T. Williams, 1980: *Numerical Prediction and Dynamic Meteorology*. 2nd ed., John Wiley & Sons, 477 pp.
- Holton, James R., 1992: *An Introduction to Dynamic Meteorology*. Academic Press Inc., 433-474.
- Hosmer, David W., and S. Lemeshow, 1989: *Applied Logistic Regression*. John Wiley & Sons, 307 pp.
- Howell, Cindy L., 1998: *Nowcasting Thunderstorms at Cape Canaveral, Florida, Using an Improved Neumann-Pfeffer Thunderstorm Index*. Masters Thesis, Air Force Institute of Technology, Wright Patterson Air Force Base, Ohio, 93 pp.
- Kelly, J. L., 1998: *Predicting East Coast Sea Breeze Initiated Convection Near Cape Canaveral, Florida*. Masters Thesis, Florida State University, 50 pp.
- Lopez, Raul E., P. T. Gannon, D. O. Blanchard, C. C. Balch, 1984: Synoptic and Regional Circulation Parameters Associated with the degree of Convective Shower Activity in South Florida. *Mon. Wea. Rev.*, 112, 686-703.
- Lopez, R. E. and R. L. Holle, 1987a: The distribution of summertime lightning as a function of low-level wind flow in Central Florida. *NOAA Technical Memorandum ERL ESG-28*.
- Manobianco, J., and Paul A. Nutter, 1998: *Evaluation of the 29-km Eta Model. Part I: Objective Verification at Three Selected Stations*. Applied Meteorology Unit, 38 pp.
- Manobianco, J., and Paul A. Nutter, 1998: *Evaluation of the 29-km Eta Model. Part II: Subjective Verification over Florida*. Applied Meteorology Unit, 50 pp.
- Mickey, J., and S. Greenland, 1989: A Study of the Impact of Confounder-Selection Criteria on Effect Estimation. *American Journal of Epidemiology*, 129, 125-137.
- NASA Contract Report. CR-205409, 1997: *Evaluation of the 29-km Eta Model for Weather Support to the United States Space Program*. Applied Meteorology Unit, 91 pp.
- NASA Contract Report. CR-1998-207910, 1998: *An Extended Objective Evaluation of the 29-km Eta Model for Weather Support to the United States Space Program*. Applied Meteorology Unit, 64 pp.
- Neter, John, W. Wasserman, M. Kutner, 1989: *Applied Linear Regression Models*. 2nd ed., Irwin, 601-616.

- Neumann, C. J., 1968: Weather Bureau Technical Memorandum SOS-2, *Frequency and Duration of Thunderstorms at Cape Kennedy, Part I*. Spaceflight Meteorology Group, 26 pp.
- Neumann, C. J., 1970: ESSA Technical Memorandum WBTM SOS-6, *Frequency and Duration of Thunderstorms at Cape Kennedy, Part II*. Spaceflight Meteorology Group, 32 pp.
- Neumann, C. J., 1971: NOAA Technical Memorandum NWS SOS-8, *Thunderstorm Forecasting at Cape Kennedy, Florida, Utilizing Multiple Regression Techniques*. Spaceflight Meteorology Group, 45 pp.
- Orszag, S. A., 1970: Transform Method for the calculation of Vector-Coupled SAMS: Application to the spectral form of the Vorticity Equation. *J. Atm. Sci.*, 27, 890-895.
- Panofsky, Hans A., and Glenn W. Brier, 1968: *Some Applications of Statistics to Meteorology*. Pennsylvania State University. 224 pp.
- Ray, P. S. ed, 1986: *Mesoscale Meteorology and Forecasting*. American Meteorological Society, 74-75, 291, 573-595, 720-751.
- Richardson, L. F., 1922: *Weather Prediction by Numerical Process*. Cambridge University Press. 236 pp.
- Roeder, William, P., 1998: Chief of Operations support Flight and Science and Technical and Training Officer for the 45th Weather Squadron, PAFB, FL. Personal E-mail's over the course of thesis preparation.
- Rogers, Eric, 1998: Environmental Modeling Center of the National Centers for Environmental Prediction. Personal E-mails over the course of thesis preparation.
- Staudenmaier, M., 1996: The Initialization Procedure in the Meso- Eta Model. *NOAA Western Region Technical Attachment #96-30*. Nov 26, 1996.
- United States Weather Bureau, 1952: Mean Number of Thunderstorm Days in the United States. Technical Paper #19, Washington, D.C.
- Wallace, John M., and Peter Hobbs, 1977: *Atmospheric Science: An Introductory Survey*. Academic Press, Inc., 85-86, 228.
- Wilks, Daniel S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, Inc., 467 pp.

- Wohlwend, C. S., 1998: *Improving Cape Canaveral's Day-2 Thunderstorm Forecasting Using Meso-Eta Numerical Model Output*. Masters Thesis, Air Force Institute of Technology, Wright Patterson Air Force Base, Ohio, 148 pp.
- Zhong, S., and Eugene Takle, 1992: An Observational Study of Sea- and Land-Breeze Circulation in an Area of Complex Coastal Heating. *Journal App. Met.*, **31**, 1426-1438.
- Zhong, S., and Eugene Takle, 1993: The Effects of Large-Scale Winds on the Sea-Land-Breeze Circulations in an area of Complex Coastal Heating. *Journal App. Met.*, **32**, 1181-1194.

Vita

John C. Crane was born on August 3, 1964 in South Ruislip Naval Hospital, United Kingdom where his father was stationed as an American liaison for the U.S. Navy. His father retired from the U.S. Navy in 1979 and moved to Berkeley Springs, West Virginia. John graduated high school in 1982 and joined the Air Force in 1985. His first station was at Patrick A.F.B., Florida where he was an enlisted aerospace ground equipment mechanic. In 1991, he was accepted into the Airman Education and Commissioning Program and was sent to the University of Oklahoma to receive his bachelor's degree in meteorology. Following his first tour as an officer at Pope A.F.B., North Carolina, he was accepted to the Air Force Institute of Technology as a master's degree student in meteorology. After completing his masters degree he will be re-assigned to Asheville, North Carolina to work at the Air Force Combat Climatology Center.

John is unmarried and has a son Drake who is 12 years old.

Permanent Address:

204 Jefferson Street

Berkeley Springs, WV 25411

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.				
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE Mar 99		3. REPORT TYPE AND DATES COVERED Masters Thesis
4. TITLE AND SUBTITLE IMPROVING CAPE CANAVERAL'S NEXT-DAY THUNDERSTORM FORECASTING USING A MESO-ETA MODEL-BASED INDEX			5. FUNDING NUMBERS	
6. AUTHOR(S) John C. Crane, 1Lt, USAF				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) AFIT/ENP 2950 P. Street Wright-Patterson AFB, OH 45433 Attn: Lt Col Michael Walters COM: (937)-255-3636- x4681 DSN: 785-3636 mwalters@afit.af.mil			8. PERFORMING ORGANIZATION REPORT NUMBER AFIT/GM/ENP/99M-4	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) 45 WS/SYR 1201 Minuteman Street Patrick AFB, FL 32925-3238 Attn: Mr William Roeder COM: (407)-852-8410 DSN: 467-8410 william.roeder@pafb.af.mil			10. SPONSORING/MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES Advisor: Michael K. Walters, Lt Col, USAF				
12a. DISTRIBUTION AVAILABILITY STATEMENT Approved for public release; distribution unlimited.			12b. DISTRIBUTION CODE A	
13. ABSTRACT (Maximum 200 words) <p>Reliable thunderstorm forecasts are essential to safety and resource protection at Cape Canaveral. Current methods of forecasting day-2 thunderstorms provide little improvement over forecasting by persistence alone and are therefore in need of replacement. This research focused on using the mesoscale eta model to develop an index for improved forecasting of day-2 thunderstorms.</p> <p>Logistic regression techniques were used to regress the occurrence of a thunderstorm at Cape Canaveral against day-2 forecast variables output, or derived, from the mesoscale eta model. Accuracy and bias scores were calculated for the forecasts made by the regression equations, and the forecast results were compared to persistence and to model-based forecasts of the Neumann-Pfeffer Thunderstorm Index (NPTI). For cases where the results were shown to be statistically significant, the forecasts made using the logistic regression equations (called the Eta Thunderstorm Index (ETI)) consistently outperformed both persistence and the NPTI. Due to the small sample size used in this research, further study on this topic is encouraged.</p>				
14. SUBJECT TERMS Meteorology, Mesoscal-eta model, Thunderstorm forecasting, Numerical Weather Prediction			15. NUMBER OF PAGES 97	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UL	